2

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b. RESTRICTIVE MARKINGS None
2a. SECURITY CLASSIFICATION AUTHORITY ---	3. DISTRIBUTION/AVAILABILITY OF REPORT Unrestricted
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE ---	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) ---	5. MONITORING ORGANIZATION REPORT NUMBER(S) AFOSR-TR- 89-1677
6a. NAME OF PERFORMING ORGANIZATION Univ. of Wisconsin-Milwaukee	6b. OFFICE SYMBOL (If applicable)
7a. NAME OF MONITORING ORGANIZATION AFOSR/NL	
6c. ADDRESS (City, State, and ZIP Code) Department of Psychology Milwaukee, WI 53201	7b. ADDRESS (City, State, and ZIP Code) Building 410 Bolling Air Force Base, DC 20332-6448
8a. NAME OF FUNDING/SPONSORING ORGANIZATION AFOSR/NL	8b. OFFICE SYMBOL (If applicable)
9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER AFOSR-88-0320	
8c. ADDRESS (City, State, and ZIP Code) Building 410 Bolling Air Force Base, DC 20332-6448	10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO.
61102F	2313	A6	

11. TITLE (Include Security Classification)
Perception of Long-Period Complex Sounds (UNCLASSIFIED)

12. PERSONAL AUTHOR(S)
Warren, Richard M.

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Annual Progress Report	FROM ~~[illegible]~~ 88 TO 31Oct89	89/11/27	

16. SUPPLEMENTARY NOTATION
1 Sep
from 19)

17. COSATI CODES		
FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
auditory perception; complex sounds; pitch

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

Working with recycled sequences of ten 40 ms items the investigators studied the discrimination of minimal changes for: ~~(1)~~ sinusoids, ~~(2)~~ vowels, and ~~(3)~~ frozen noise segments. Listeners made ABX judgments for sequences differing only in the ordering of two contiguous items. In contrast with results previously obtained for ten-item sequences presented in transient "one-shot" bursts, recycled stimuli were readily discriminated by untrained listeners. The relative difficulty of discriminating tonal patterns (measured by response time) was an inverse function of: a) the frequency separation between the permuted tones; and b) the frequency separation between the tones immediately preceding and following the permuted pair. For the vowel sequences, listeners' trial by trial repor indicate that discrimination of order was mediated by verbal organization involving introduction of illusory consonants and distortion of the vowels. Discrimination of order within sequences of frozen noise was more difficult than found with tone or vowel sequence but all listeners performed at levels well above chance. Additional work with recycled›

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED □ SAME AS RPT. □ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION UNCLASSIFIED
22a. NAME OF RESPONSIBLE INDIVIDUAL John F. Tangney, Program Manager	22b. TELEPHONE (Include Area Code) (202) 767-
22c. OFFICE SYMBOL AFOSR/NL	

DD Form 1473, JUN 86 Previous editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE

Block #19 cont'd

frozen noise is proceeding satisfactorily which deals with the ability to remember and recognize segments up to 1 s in duration, and the relative salience of various spectral regions in this process.

Keywords:

[illegible]

Report AFOSR-88-0320

PERCEPTION OF LONG-PERIOD COMPLEX SOUNDS

Richard M. Warren
University of Wisconsin-Milwaukee
Department of Psychology
Milwaukee, Wisconsin 53201

27 November 1989

Annual Progress Report for Period 1 September 1988 - 31 October 1989

Prepared for
AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
Building 410
Bolling Air Force Base, DC 20332-6448

Richard M. Warren
AFOSR Grant No. 88-0320

Annual Progress Report AFOSR

SUMMARY

Working with recycled sequences of ten 40 ms items the investigators studied the discrimination of minimal changes for: (1) sinusoids, (2) vowels, and (3) frozen noise segments. Listeners made ABX judgments for sequences differing only in the ordering of two contiguous items. In contrast with results previously obtained for ten-item sequences presented in transient "one-shot" bursts, recycled stimuli were readily discriminated by untrained listeners. The relative difficulty of discriminating tonal patterns (measured by response time) was an inverse function of: a) the frequency separation between the permuted tones; and b) the frequency separation between the tones immediately preceding and following the permuted pair. For the vowel sequences, listeners' trial by trial reports indicate that discrimination of order was mediated by verbal organization involving introduction of illusory consonants and distortion of the vowels. Discrimination of order within sequences of frozen noise was more difficult than found with tone or vowel sequences, but all listeners performed at levels well above chance. Additional work with recycled frozen noise is proceeding satisfactorily which deals with the ability to remember and recognize segments up to 1 s in duration, and the relative salience of various spectral regions in this process.

STATEMENT OF WORK

The objective of the AFOSR supported research is to further our knowledge of mechanisms and principles governing the perception of complex sounds. The major research completed to date deals with the perception of repeated sequences. This work suggests that sequences consisting of brief sounds (item durations less than 100 ms) are not perceived as a succession of discrete items, but rather as overall patterns without resolution into component sounds.

Watson and his coworkers have employed sequences of ten or more brief sinusoidal tones ("Watson sequences") in studies examining the ability to make fine discriminations within complex "word-length" patterns. These studies, which employed contrasting sequences presented in single statements, have shown that listeners usually require many hours of training before discrimination of tones embedded within sequence bursts approximates discrimination involving the same tones presented in isolation (see Watson, 1987 for a review).

In contrast with the extensive training required for discriminating between Watson sequences presented as "one-shot" transient patterns, discrimination involving simpler sequences consisting of three or four sounds generally can be performed by untrained listeners within one minute of A/B comparison when the sequences are recycled (e.g., Warren & Obusek, 1972). The present investigator reasoned that discrimination of Watson sequences might be greatly facilitated if these stimuli were presented as recycling patterns, and this hypothesis was supported by preliminary listening. When permitted to switch at will between recycling 10-item sequences of 40 ms tones, it was found that listeners required only a few tens of seconds to detect a difference in the frequency, intensity or duration of a single component. This ease of discrimination held even under conditions of high uncertainty (concerning the dimension subject to change), and seemed to involve "hearing out" the altered component as an intermittent, isolated tone. Because this analytic process enabled listeners to rapidly achieve errorless discrimination of changes in the identity (e.g, frequency) of components, the investigator's formal experiments with complex sequences examined the more difficult task of discriminating minimal changes in temporal order.

Listeners made ABX judgments for pairs of recycling sequences of ten 40 ms items that differed in the ordering of two contiguous sounds. In addition to sequences of sinusoidal tones (Experiment 1), listeners were also presented with sequences of monotone vowels (Experiment 2) and with sequences of frozen noise segments (Experiment 3). The vowel and noise stimuli were included to permit an examination of the effect of recycling upon order discrimination within broadband patterns of two rather different types: one type consisting of discrete, identifiable components, and the other type consisting of novel waveforms not readily identified as a sequence of discrete items.

The investigators employed recycling patterns consisting of ten 40 ms components. Components of all patterns were presented at equal amplitude (70 dB) and had linear rise/fall times of 2.5 ms. A pattern pool of forty-eight tonal sequences was constructed by sampling from a catalog of 10 sinusoidal frequencies, ranging from 500 to 1500 Hz in equilog steps (This pool was similar to that used by Watson et al., 1975, experiment 4). Next, forty-eight vowel sequences were constructed from a catalog of ten vowels (The vowels were those in: 'heed', 'hid', 'head', 'had', 'hod', 'hawd', 'hood', 'hud', 'hoot' and 'herd'). Each 40 ms vowel segment consisted of 8 iterations of a single 5 ms glottal pulse, originally excised in digital form from a sustained monotone production (200 Hz voicing frequency). Finally, forty-eight noise sequences were constructed by sampling from a catalog of 10 random waveforms that were originally excised from on-line white noise. Vowel and noise sequences were bandpass filtered from 50 Hz to 8000 Hz with slopes of 48 dB/octave.

Contrasting "A" and "B" versions of the 48 sequences in each stimulus category were created by permuting the order of a single pair of contiguous items.

The four subjects participating in the study were tested individually in an audiometric room with stimuli delivered through headphones at 70 dB SPL. They were provided with a three-button panel which they used for switching between contrasting "A" and "B" stimuli and a third "X" stimulus which was identical to the sequence presented in either the A or B channel. Listeners switched at will between the three signals until satisfied that they had determined a match. They were aware that their responses were being timed, and they received trial-by-trial feedback concerning their matching accuracy.

Listeners were presented first with all tonal sequences, then with the vowel sequences, and finally the noise sequences. The forty-eight contrasting pairs of sequences in each category were presented twice to each listener for a total of 96 judgments for each type of sequence. Each listener received a different random ordering of stimuli for their first judgment of each contrasting pair of sequences, and this order was repeated for that listener upon second presentation of the stimuli, so that the two judgments for each contrast were separated by judgments for the remaining 47 sequence pairs.

The number of correct responses and the median response times for judgments are presented in Tables 1-3. As shown, overall matching accuracy was well above chance for all listeners and for each type of pattern. The percentage of correct responses ranged from about 90% to 98% for tonal sequences, from 96% to 98% for vowel sequences, and from 80% to 98% correct for sequences of frozen noise. For the tonal sequences presented in the first experiment, judgments made by the experienced listeners BB and JB were more rapid and somewhat more accurate ($Z = 2.38$, $p < .02$) than those of the naive listeners JR and KR. However, all listeners performed at above chance levels ($p < .05$ or better) across their two blocks of trials (12 judgments total) for both tone and vowel sequences, and the three listeners who participated in Experiment 3 scored above chance ($p < .05$ or better) for their initial 12 judgments with sequences of frozen noise.

Table 1: Accuracy and Response Times for ABX Judgments of Recycled Ten-Tone Sequences (A and B sequences differed in the order of a single, contiguous pair of tones.)

	Number Correct out of 96*	Response Times (SEC)		
Listener		Median	Q1	Q3
BB	94	25.0	15.5	37.5
JB	94	21.5	15.0	33.5
JR	86	89.5	45.0	161.5
KR	87	66.0	40.5	119.0

* Accuracy scores for all listeners exceeded chance ($z > 3.46$, $p < .001$)

Table 2: Accuracy and Response Times for ABX Judgments of Recycled Ten-Vowel Sequences (A and B sequences differed in the order of a single, contiguous pair of vowels.)

Listener	Number Correct out of 96*	Response Times (SEC)		
		Median	Q1	Q3
BB	92	34.5	25.0	51.0
JB	94	50.5	30.0	107.5
JR	94	72.0	41.0	114.0
KR	94	42.0	28.0	68.5

* Accuracy scores for all listeners exceeded chance (z > 3.46, p < .001)

Table 3: Accuracy and Response Times for ABX Judgments of Recycled Ten-item Sequences of Frozen Noise Segments (A and B sequuences differed in the order of a single, contiguous pair of noise segments.)

	Number Correct out of 96*	Response Times (SEC)		
Listener		Median	Q1	Q3
BB	94	71.5	41.5	123.0
JB	94	55.5	37.0	93.5
JR	77	193.0	89.0	340.5
KR	84	111.0	58.0	252.0

* Accuracy scores for all listeners exceeded chance ($z > 3.46$, $p < .001$)

The results overall indicate clearly that discrimination of order within word-length sequences is readily accomplished when the patterns are repeated without pause. This ease of discrimination applies not only for sequences comprised of sounds differing systematically in pitch (sinusoids) or identifiable quality (vowels), but also for sequences of complex, arbitrary components lacking discrete identities. Other aspects of the results for specific types of sequences are discussed in greater detail below.

Tonal Sequences

The accuracy and response times for judgments of ten-tone sequences were comparable to those previously reported for order discrimination within simpler recycled sequences of 3 or 4 items (Warren & Obusek, 1972). Correlational analysis of response times for judgments of tonal sequences showed substantial consistency in the relative difficulty of specific patterns both within listeners ($R = .4$ to $.66$, $p < .05$ or better) and across listeners ($R = .35$ to $.67$, $p < .02$ or better). It was also found that approximately 70% of the variance in response times for tonal sequences could be accounted for by two factors (Multiple $R = 0.84$, $p < .0001$): Times for judgments decreased with increasing frequency separation of the permuted tones, and, also decreased with increasing frequency separation of the tones immediately preceding and following the permuted pair. As these frequency separations increased, there was a tendency for each of the permuted tones to be grouped perceptually with an adjacent nonpermuted tone in only one order, resulting in what was typically described as a difference in the "rhythmic complexity" of the contrasting sequences. Somewhat similar results have been reported by Nickerson and Freeman (1974) for recycled 4-item sequences of 200 ms tones.

Vowel Sequences

Accuracy and response times for judgments of vowel sequences were similar to those obtained for sequences of pure tones, and, as was found for sinusoids, there was substantial consistency in response times for specific vowel patterns both within listeners ($R = .36$ to $.66$, $p < .05$ or better) and across listeners ($R = .30$ to $.46$, $p < .05$ or better). However, listeners' trial by trial reports concerning the nature of their discriminations indicate that rather different processes were involved for tonal and vowel stimuli. Listeners typically reported that contrasting tonal sequences differed in rhythmic complexity, whereas contrasting vowel sequences often evoked different compelling verbal organizations—occasionally corresponding to pseudowords but most often real words. Thus, for 57% of the trials with vowel sequences, listeners reported using differences in verbal organization as the basis of their judgments. In most cases, listeners reported that only one of the contrasting sequences was organized verbally, but in some cases different words were reported for the two orderings (e.g., "valuable" vs "technical"). Most interestingly, although there was little agreement across listeners in the verbal forms evoked by specific vowel sequences, there was substantial consistency within listeners: In 52% of the cases (50 out of 96) in which listeners reported specific words upon first presentation of sequences, they reported the same word or words on second presentation--this was in spite of the fact that successive judgments of the same sequence were separated by several days and by interpolated judgments of the remaining 47 contrasting sequences. Thus, although "verbal summation" of these monotone vowel patterns was highly idiosyncratic, it was also remarkably stable.

Frozen Noise Sequences

Not surprisingly, discriminating minimal changes in order within the complex, novel patterns formed by concatenated segments of frozen noise required more time and, for two listeners, was less accurate than discrimination of tone or vowel sequences ($p < .01$ or better). Listeners readily detected recycling of the 400 ms noise sequences: They reported hearing the "whooshing" described by Guttman and Julesz (1963) for iterated frozen noise segments of this duration, as well as a variety of repetitive transient sounds such as "bumps", "beeps", and "clanks". Not all organizations initially "heard out" by listeners were altered perceptibly by minimal permutation of noise segments. However, listeners did find that new organizations continued to emerge with extended listening to a contrasting pair of sequences, and in most cases a pattern could be found that varied with permutation--typically either in quality or rhythmic complexity. The great variety and instability of perceptual organizations heard with a single recycled sequence may have been responsible for the fact that response times for specific sequences of noise were not correlated across listeners.

Additional work is in progress which measures the ability of listeners to recognize repeated frozen noises after delays exceeding the limits of short term "echoic" storage. In addition, the investigator is comparing the salience of various spectral regions in the recognition of frozen noise segments.

REFERENCES

Guttman, N. & Julesz, B. (1963). Lower limits of auditory periodicity analysis. Journal of the Acoustical Society of America, 35, 610 (L).

Nickerson, R.S., & Freeman, B. (1974). Discrimination of the order of the components of repeating tone sequences: Effects of frequency separation and extensive practice. Perception & Psychophysics, 16, 471-477.

Warren, R.M., & Obusek, C.J., (1972). Identification of temporal order within auditory sequences. Perception & Psychophysics, 12, 86-90.

Watson, C.S., (1987). Uncertainty, informational masking, and the capacity of immediate memory. In W.A. Yost and C.S. Watson (eds.), Auditory Processing of Complex Sounds. New Jersey: Lawrence Erlbaum Associates, pp. 267-277.

Watson, C.S., Wroton, H.W. Kelly, W.J., & Benbassat, C.A. (1975). Factors in the discrimination of tonal patterns. I. Component frequency, temporal position, and silent intervals. Journal of the Acoustical Society of America, 57, 1175-1185.

PUBLICATIONS SINCE PREVIOUS REPORT

1. Warren, R.M. "Perceptual bases for the evolution of speech." In M. Landsberg (Ed.), Genesis of Language. Berlin: Gruyter, 1988, pp. 101-110.

2. Warren, R.M. & Bashford, J.A. Jr. "Broadband repetition pitch: Spectral dominance or pitch averaging?" Journal of the Acoustical Society of America, 1988, 84, 2058-2062.

3. Warren. R.M., Wrightson, J.M., & Puretz, J. "Illusory continuity of tonal and infratonal periodic sounds." Journal of the Acoustical Society of America, 1988, 84, 1338-1342.

4. Bashford, J.A. Jr., Meyers, M.D., Brubaker, B.S., & Warren, R.M. "Illusory continuity of interrupted speech: Speech rate determines durational limits." Journal of the Acoustical Society of America, 1988, 84, 1635-1638.

5. Warren, R.M., Bashford, J.A. Jr., & Brubaker, B.S. "Perception of complex tones mistuned from unison." Journal of the Acoustical Society of America, 1989, 86, 116-125.

6. Warren, R.M., Gardner, D.A., Brubaker, B.S. & Bashford, J.A. Jr. "Melodic and nonmelodic pitch patterns: Effects of duration on perception." Proceedings of the First International Congress on Music Perception and Cognition, Kyoto, Japan, October 1989, pp. 343-348.

7. Warren, R.M. "Sensory magnitudes and their physical correlates." (Commentary on Target Article "Reconciling Fechner and Stevens: Toward a unified psychophysical law" by L.E. Krueger). Behavioral and Brain Sciences, 1989, 12, 296-297.

8. Warren, R.M. (One of 9 co-authors of the Report of the CHABA Panel on Classification of Complex Nonspeech sounds, W.A. Yost, Chair). National Academic Press, 1989, 88 pp.

PUBLICATION IN PRESS

1. Warren, R.M. & Bashford, J.A. Jr., "Tweaking the lexicon: Organization of vowel sequences into words." To be published in Perception & Psychophysics.

PROFESSIONAL PERSONNEL

In addition to R.M. Warren, James A. Bashford, Jr., Ph.D. is participating in the project in the capacity of Associate Researcher. Graduate students who have been assisting in the project this past year are Bradley S. Brubaker, Keri R. Reiner, Jill Robertson, and Daniel G. Zuck.

PROFESSIONAL INTERACTIONS

1. Papers presented at 116th Acoustical Society of America Meeting, November, 1988.

   a. "Discrimination of recycled word-length sequences", Journal of the Acoustical Society of America, 1988, 84, S141 (Abstract).

   b. "Learning to identify phonemic orders", Journal of the Acoustical Society of America, 1988, 84, S154 (Abstract).

2. "Melodic and Nonmelodic pitch patterns: Effects of duration on perception," Invited paper presented at the First International Conference on Music Perception and Cognition, Kyoto, Japan, October, 1989, (also served as session chairman). Paper published in the Conference Proceedings, pp. 343-348.

3. Invited participant at Osaka Symposium on Perception, sponsored by the Department of Psychology, University of Osaka, Japan, October 1989.

APPENDICES

A. Warren, R.M. "Perceptual bases for the evolution of speech." In M. Landsberg (Ed.), Genesis of Language. Berlin: Gruyter, 1988, pp. 101-110.

B. Warren, R.M. & Bashford, J.A. Jr. "Broadband repetition pitch: Spectral dominance or pitch averaging?" Journal of the Acoustical Society of America, 1988, 84, 2058-2062.

C. Warren. R.M., Wrightson, J.M., & Puretz, J. "Illusory continuity of tonal and infratonal periodic sounds." Journal of the Acoustical Society of America, 1988, 84, 1338-1342.

D. Bashford, J.A. Jr., Meyers, M.D., Brubaker, B.S., & Warren, R.M. "Illusory continuity of interrupted speech: Speech rate determines durational limits." Journal of the Acoustical Society of America, 1988, 84, 1635-1638.

E. Warren, R.M., Bashford, J.A. Jr., & Brubaker, B.S. "Perception of complex tones mistuned from unison." Journal of the Acoustical Society of America, 1989, 86, 116-125.

F. Warren, R.M., Gardner, D.A., Brubaker, B.S. & Bashford, J.A. Jr. "Melodic and nonmelodic pitch patterns: Effects of duration on perception." Proceedings of the First International Congress on Music Perception and Cognition, Kyoto, Japan, October 1989, pp. 343-348.

G. Warren, R.M. & Bashford, J.A. Jr., "Tweaking the lexicon: Organization of vowel sequences into words." Manuscript of paper to be published in Perception & Psychophysics.

Studies in Anthropological Linguistics

3

Editors
Florian Coulmas
Jacob L. Mey

Mouton de Gruyter
Berlin · New York · Amsterdam

Appendix A

The Genesis of Language

A Different Judgement of Evidence

edited by
Marge E. Landsberg

Mouton de Gruyter
Berlin · New York · Amsterdam 1988

Clark, R. A.
1978 The transition from action to gesture. In A. J. Lock (ed.), *Action, Gesture and Symbol: The Emergence of Language*. 231–257. London: Academic Press.

Fenton, M. B. and J. H. Fullard
1981 Moth hearing and the feeding strategies of bats. *American Scientist* 69. 266–275.

Hewes, G. W.
1980 Comment. *Current Anthropology* 21. 781–782.
1984 The invention of phonemically-based language. In A. J. Lock and E. Fisher (eds.), *Language Development*. 49–57. London: Croom Helm.

Krantz, G.
1980 Sapientization and speech. *Current Anthropology* 21. 773–792.

Laitman, J. T.
1983 The evolution of the hominid upper respiratory system and implications for the origins of speech. In Eric de Grolier (ed.), *Glossogenetics*. The origin and evolution of language. Proceedings of the International Transdisciplinary Symposium on Glossogenetics. 63–90. Chur, Switzerland: Harwood Academic Publishers.

Lock, A. J.
1980 *The Guided Reinvention of Language*. London: Academic Press.
1981 The early stages of communicative and linguistic development: Underlying processes. In B. de Gelder (ed.), *Knowledge and Representation*. 94–110. London: Routledge and Kegan Paul.
(in pr.) Underlying processes in the elaboration of language. In E. S. Gollin (ed.), *The Evolution of Adaptive Behavior*. Hillsdale, NJ: Erlbaum.

Mead, G. H.
1934 *Mind, Self and Society*. Chicago: Chicago University Press.

Plooij, F. X.
1978 Some basic traits of language in wild chimpanzees? In A. J. Lock (ed.), *Action, Gesture and Symbol: The Emergence of Language*. 111–131. London: Academic Press.

Reynolds, P.
1983 Ape constructional ability and the origin of linguistic structure. In Eric de Grolier (ed.), *Glossogenetics*. The origin and evolution of language. Proceedings of the International Transdisciplinary Symposium on Glossogenetics. 185–200. Chur, Switzerland: Harwood Academic Publishers.

Terrace, H. S.
1979 *Nim: A Chimpanzee Who Learned Sign Language*. New York: Knopff. (Cited by Reynolds, 1983).

Volosinov, V. N.
1973[1929] *Marxism and the Philosophy of Language*. New York: Seminar Press.

Wynn, T.
1979 The intelligence of later Acheulian hominids. *Man* 14. 371–391.

Wynn, T.
1981 The intelligence of Oldowan hominids. *Journal of Human Evolution* 10. 529–431.

Perceptual bases for the evolution of speech

Richard M. Warren

Abstract

It is suggested that speech perception is based upon a holistic recognition of complex acoustic patterns, and does not require the ability to identify individual component sounds. Much confusion in the literature is associated with attempts to consider that speech perception requires the ability to recognize phonemes and their orders at some level of perceptual organization. There is evidence that our ability to recognize acoustic patterns holistically is shared with other animals, and that speech perception evolved from this prelinguistic ability. It appears that identification of component sounds and their orders is a linguistic skill which is the consequence of, not the basis of, speech recognition.

Introduction

Before we can begin to trace the evolution of speech and language, it is necessary to understand the nature of mechanisms used for speech perception. Unfortunately, a pervasive emphasis upon phonemes as linguistic units has impaired our understanding the nature of speech perception and its development from auditory capabilities of our prelinguistic ancestors. This paper will attempt to demonstrate that:

1. The concept of phonemes as units of speech can be traced back to the invention of the alphabet.
2. The term 'phoneme' as used today has multiple meanings (articulatory, acoustic, perceptual, and graphemic), and the use of the same term for different entities has led to considerable confusion along with inappropriate theories of speech perception.
3. Sound patterns consisting of sequences of several acoustic 'phonemes' serve as units of organization in speech perception.

4. Animals other than man are capable of differentiating between complex acoustic sequences, including those of speech.
5. The emphasis placed by some theorists upon the lack of speech-producing capabilities of nonhuman primates and other animals may not be directly relevant to an understanding of the differences in linguistic capacity between humans and other creatures.
6. While human languages have evolved as a form of acoustic communication, these languages can readily be extended into non-phonetic acoustic modes, as well as a number of forms employing sensory modalities other than hearing. A cross-modality comparison of the modes of linguistic communication should be useful in understanding the essential characteristics of human language.

Multiple meanings of the term 'phoneme'

The use of the same term to describe different entities can impair the development of a science. In a recent paper (Warren, 1983) I have attempted to show that there are four different uses of the term 'phoneme': a) The articulatory phoneme refers to units employed in the production of speech; b) the acoustic phoneme refers to units employed to classify the sounds of speech; c) the perceptual phoneme refers to units employed in the auditory organization of heard speech; d) the graphemic phoneme refers to the written symbol employed to designate any or all of the other three classes of phonemes. As we *shall* see, the lack of correspondence between entities bearing the same name has caused great confusion concerning the nature of speech, and this confusion has implications for theories concerning the evolution of speech.

The alphabet and its relation to graphemic and articulatory phonemes

The concept that speech can be analyzed into a sequence of phonemes can be traced back to alphabetic writing (for discussion, see Warren, 1983). Unlike other fotms of writing, the alphabet seems to have been invented only once, and to have spread rapidly to other cultures. The alphabet is based upon articulatory activities employed in generating speech. It was an insight of considerable

utility to consider that there are a limited number of ways of producing sounds used in a particular language, and that by using a separate written symbol for each of these sound-generating activities, it is possible to transcribe speech as a sequence of articulatory gestures. Note that I have described this alphabetic analysis of speech in terms of articulatory activities *rather* than sounds (the evidence for and the significance of this distinction will emerge shortly). It is possible to analyze and tabulate these activities readily by direct observation involving oneself and others. The positions employed for consonants are in general easiest to observe, and historically consonants were transcribed by graphemes first. The manner of producing vowels is not as readily observable, and early alphabetic writing did *not* include symbols for vowels. Writing with a full alphabet of consonants plus vowels, permits an unfamiliar word to be pronounced, since the string of graphemes not only represents the word but provides instructions for its production. Of course, the graphemes used for languages such as English may diverge considerably from current pronunciation. However, it is still possible for readers to pronounce many unfamiliar printed English words with some degree of accuracy. Other languages have maintained closer correspondence between orthography and pronunciation than English and, as we know, the 'phonetic' alphabet employs a series of graphemes designed especially to correspond closely to spoken language.

Differences between articulatory phonemes and acoustic phonemes (speech sounds)

It is often assumed that every articulatory phoneme has a corresponding acoustic phoneme. However, devices capable of analyzing speech sounds acoustically have indicated that this assumption is false. The lack of correspondence between articulatory phonemes and their acoustic consequences has resulted in what Klatt (1979) has called the "acoustic-phonetic non-invariance problem." To take one example, the acoustical nature of the articulatory phoneme /d/ in /di/ is quite different from the acoustical nature of the /d/ in /du/ (see Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967). The great effects of neighboring speech sounds upon the nature of acoustic 'phonemes' are evident when attempts are made to read sound spectrograms which display the results of a spectral analysis in visual form. The sound spectrograph was developed in the 1940s by the Bell Laboratories in the hope of enabling the deaf to understand speech through vision (Potter, Kopp and Kopp, 1947). However, even with considerable practice, it is not possible to use such a display for real-time perception of speech, due to the varied acoustic forms of the same 'phonemes.' Nevertheless, there are those who maintain that although some acoustic characteristics of a speech sound change with context, there may be other invariant cues (not readily apparent through acoustic analysis), which are used by listeners for identification of acoustic phonemes (see Jusczyk, Smith and Murphy, 1981; Stevens and Blumstein, 1981).

Is there a perceptual phoneme?

Most theories of speech perception have assumed the existence of phonetic units at some level of auditory analysis (for discussion, see Warren, 1982, 1983). However, there is now considerable evidence that phonetic analysis is not necessary for speech perception, and probably does not take place as a precursor to comprehension. Many of the persistent and ingenious attempts to demonstrate the invariance of acoustic phonemes result from the need to use such entities for phonetically-based perceptual theories. But if there are no phonetic perceptual units, then this need vanishes.

Let us examine some of the evidence that phonemes are not units for the perception of speech. It has been shown that before children can read, they have great difficulty in segmenting words into speech sounds corresponding to phonemes or graphemes (Calfee, Chapman and Venezky, 1972; Gibson and Levin, 1975; Gleitman and Rozin, 1973; Savin, 1972). Once children have progressed to reading in school, then division of words into phonemes becomes possible (Liberman, Shankweiler, Fischer and Carter, 1974). It might be considered that the facilitation of phonetic segmentation results from developmental changes and increased linguistic skills rather than the acquisition of reading ability. However, Morais, Bertelson, Cary and Alegria (1986) reported that adults who had never learned to read could not recognize, delete, or add phonemes to words, while other members of the same population of illiterates could perform these tasks involving phonemes following training in special adult reading classes.

Another line of evidence indicating that phonemes are not employed as perceptual units is provided by reaction time studies. It has been shown that the time required to react to phoneme targets in syllables is greater than the time required to react to the syllables themselves (Savin, and Bever, 1970). These results suggested to Savin and Bever that the phoneme may be derived from prior identification of the syllable, rather than serving as the unit requiring identification before the syllable can be recognized. Support for this view was afforded in a study by Warren (1971) in which prior syntactic and semantic contexts within sentences were manipulated to vary the probability of occurrence of target words. As anticipated, a more likely word was identified more quickly. However, the point of interest for this discussion is that a contextually facilitated reaction time to a word, as measured for one group of subjects, was associated with a similar facilitation of the reaction time to an individual phoneme target within that word, as measured for a separate group of subjects. This is in keeping with the hypothesis that phonemes are derived perceptually from words, not the words from phonemes.

Consequences of nonphonetic theories of speech perception for theories of speech evolution

If we rid ourselves of the belief that speech perception rests upon special processing requiring the identification of component phonemes and their orders, then several questions suggest themselves, such as whether equivalent rules govern the perception of acoustic patterns in other animals, and whether the rules governing speech recognition also govern the recognition of nonverbal patterns in humans. A number of investigators have shown that nonhuman animals could be taught to discriminate between different isolated phonemes and between different syllables. Thus, it has been shown by Dewson (1964) that cats can learn to distinguish between the vowels "ee" and "oo" whether spoken by a woman or a man. Kuhl and Miller (1978) taught chinchillas to discriminate between the voiced and unvoiced consonant pairs represented by /kah/ and /gah/, /pah/ and /bah/, and /tah/ and /dah/. Warfield, Rubin and Glackin (1966) reported that cats could be taught to discriminate between /cat/ and /bat/, and that the limit of acoustic distortion permitting discrimination was similar for cats and humans. Since it cannot be argued that these animals appear to have evolved genetically determined mechanisms specialized for human speech sounds, these studies must be tapping some general mechanisms for detection of acoustic sequences. It has been suggested that humans and other animals possess mechanisms for perceiving complex patterns holistically, so that the pattern is recognized as an entity without the need for analysis as a sequence of identifiable items in a particular order (for a discussion, see Warren, 1982). Studies of sequences of hisses, tones, and buzzes have helped to demonstrate that we share the ability to recognize complex acoustic sequences holistically with animals, and that this ability serves as the basis for the perception of speech.

Holistic pattern recognition in humans

Several studies have demonstrated that humans can discriminate between permuted orders in otherwise identical sequences consisting of nonspeech sounds, even when the acoustic components are too brief to be identified.

Efron (1973) and Yund and Efron (1974) have found that listeners could distinguish between 'micropatterns' consisting of permuted orders of two-item sequences (for example, two tones), when the separation between the sounds was only one or two msec. Listeners appeared to discriminate on the basis of qualitative differences, and could not identify the order of components. These observations were confirmed in essential details by Wier and Green (1975).

Two-item sequences are rather special, and the use of iterated sequences of three or four sounds was introduced by me as a way of studying the perception of continuing sequences consisting of only a few items (Warren, 1968; Warren, Obusek, Farmer and Warren, 1969). It was found that three- or four-item 'recycled' sequences of nonverbal sounds require at least 200 msec/item for identification of the order of items, yet it is possible to distinguish readily between different arrangements of the same sounds down to five or ten msec/item whether subjects are trained (Warren, 1974*a*) or untrained (Warren, 1974*b*). While discriminating between permuted orders of brief items is accomplished on the basis of qualitative or holistic perceptual differences, the ability to discriminate between different orders of items having durations longer than a few hundred milliseconds appears to rest upon the linguistic skill of naming items in their appropriate order, and remembering this sequence of *names* (Warren, 1974*a*; Teranishi, 1977). We shall return to the use of verbal mechanisms for discriminating between different arrangements of long-duration items later, when we discuss sequence perception in animals other than humans. At this point, it should be noted that there is no upper limit for item durations permitting discrimination of permuted orders in humans.

Holistic pattern recognition in nonhuman mammals

A few studies have examined the ability of nonhuman mammals to distinguish between permuted orders of discrete sounds. While each of these studies has found that the animals employed could discriminate between permuted orders of sounds having brief durations, it was observed that a breakdown in the ability to distinguish between different orders of the same items occurred when the item durations exceeded more than a few seconds.

Dewson and Cowey (1969) taught monkeys to discriminate between the four possible pairs of sounds which can be generated using a tone and a hiss (*tone-hiss*, *hiss-tone*, *hiss-hiss*, *tone-tone*) when items had durations of less than about 1.5 sec. At item durations of three sec and greater, the monkeys could not perform the task, and it appeared that they were unable to remember the first item after the second item ended (they were not permitted to respond until the sequence was completed). Monkeys are primarily visual rather than auditory, and their failure to master the task at long-item durations might be attributed to a general difficulty with auditory tasks. However, a similar experiment was carried out using the dolphin (Thompson, 1976), a creature generally considered both highly intelligent and primarily auditory rather than visual in its normal activities. Four sounds, which can be designated as A, B, C, and D, were used to construct sequences of two sounds which were presented through hydrophones. The dolphin was rewarded if it pressed one paddle following the sequences AC or BD, or if it pressed a different paddle following the sequences AD or BC. The sounds had a fixed duration, and a silent period of variable length was inserted between the first and second sounds of the pairs. In order to respond appropriately, the dolphin needed to remember the first sound until the second sound occurred. Thompson reported that nearly perfect performance was obtained when the interval separating the sounds was less than two or three seconds. At longer temporal separations, performance was at chance levels. He concluded that the ability to hear the overall pattern ceased at the upper limit of behavioral discrimination, and that the perception of the overall pattern was required for a correct response.

The evidence which has been summarized suggests that speech perception is based upon the ability to recognize patterns of sounds holistically, and that we share this ability with other animals. Our perception of speech does not require the identification of component sounds and their orders – rather the identification of components and their orders within acoustic sequences is itself a linguistic skill.

What is special about human linguistic skills?

There seems little doubt that human language originated and evolved as an acoustically based method of communication employing sounds generated by our vocal tract. However, our use of language today does not require conventional speech sounds – whistled languages which remain intelligible over great distances have been developed as an ancillary method of communication in a number of mountainous areas (Busnel and Classe, 1976). Language does not even require acoustic signals: Reading is every bit as rapid and accurate in transmitting linguistic information, and sign languages are used with fluency by the deaf. Languages using visual signs need not correspond directly to a spoken language (as does signed English), but can develop into uniquely visual forms with quite different rules (as does American Sign Language). Language does not even require use of our special distance senses of hearing and vision: The sense of touch can be used by the blind-deaf in communication, and braille permits tactual reading by the blind.

Hence, although the development of special sound-producing systems seems to be associated with the evolution of human language, linguistic communication can now operate without the use of sound, when necessary. It seems that our use of language is based upon an ability to manipulate symbols according to learned conventions in an exceedingly complex and versatile fashion. These symbols can consist of auditory, visual, or tactile patterns. It is through the study of this symbol-manipulative ability within and across sensory modalities that we can more fully understand the mechanisms subserving human language and the evolutionary development of speech.

References

Busnel, R. G. and A. Classe
1976 *Whistled Languages*. New York: Springer.
Calfee, R., R. Chapman and R. Venezky
1972 How a child needs to think to learn to read. In L. W. Gregg (ed.), *Cognition in Learning and Memory*. 139–182. New York: Wiley.
Dewson, J. H. III
1964 Speech sound discrimination by cats. *Science* 144. 555–556.
Dewson, J. H. III and A. Cowey
1969 Discrimination of auditory sequences by monkeys. *Nature* 222. 695–697.
Efron, R.
1973 Conservation of temporal information by perceptual systems. *Perception and Psychophysics* 14. 518–530.
Gibson, E. J. and H. Levin
1975 *The Psychology of Reading*. Cambridge, MA: MIT.
Gleitman, L. R. and P. Rozin
1973 Teaching reading by use of a syllabary. *Reading Research Quarterly* 8. 447–483.
Jusczyk, P. W., L. B. Smith and C. Murphy
1981 The perceptual classification of speech. *Perception and Psychophysics* 30. 10–23.
Klatt, D. H.
1979 Speech perception: A model of acoustic-phonetic analysis and lexical access. *Journal of Phonetics* 7. 279–312.
Kuhl, P. and J. D. Miller
1978 Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. *Journal of the Acoustical Society of America* 63. 905–917.
Liberman, A. M., F. S. Cooper, D. P. Shankweiler and M. Studdert-Kennedy
1967 Perception of the speech code. *Psychological Review* 74. 431–461.
Liberman, I. Y., D. Shankweiler, F. W. Fischer and B. Carter
1974 Reading and the awareness of linguistic segments. *Journal of Experimental Child Psychology* 18. 201–212.

Morais, J., P. Bertelson, L. Cary and J. Alegria
1986 Literacy training and speech segmentation. *Cognition* 24, 45–64.

Potter, R. K., G. A. Kopp and H. G. Kopp
1947 *Visible Speech*. New York: Van Nostrand.

Savin, H. B.
1972 What the child knows about speech when he starts to learn to read. In J. F. Kavanagh and I. G. Mattingly (eds.), *Language by Ear and by Eye*. 319–329. Cambridge, MA: MIT.

Savin, H. B. and T. G. Bever
1970 The nonperceptual reality of the phoneme. *Journal of Verbal Learning and Verbal Behavior* 9. 295–302.

Stevens, K. N. and S. E. Blumstein
1981 The search for invariant acoustic correlates of phonetic features. In P. D. Eimas and J. L. Miller (eds.), *Perspectives on the Study of Speech*. 1–38. Hillsdale, NJ: Erlbaum.

Teranishi, R.
1977 Critical rate for identification and information capacity in hearing system. *Journal of the Acoustical Society of Japan* 33. 136–143.

Thompson, R. K. R.
1976 Performance of the bottlenose dolphin (*Tursiops truncatus*) on delayed auditory sequences and delayed auditory successive discriminations. Doctoral Dissertation, University of Hawaii.

Warfield, D., R. J. Rubin and R. Glackin
1966 Word discrimination in cats. *Journal of Auditory Research* 6. 97–119.

Warren, R. M.
1968 Relation of verbal transformation to other perceptual phenomena. Conference Publication No. 42, *Institution of Electrical Engineers* (London), Supplement No. 1. 1–8.
1971 Identification times for phonemic components of graded complexity and for spelling of speech. *Perception and Psychophysics* 9. 345–349.
1974*a* Auditory temporal discrimination by trained listeners. *Cognitive Psychology* 6. 237–256.
1974*b* Auditory pattern discrimination by untrained listeners. *Perception and Psychophysics* 15. 495–500.
1982 *Auditory Perception: A New Synthesis*. New York: Pergamon.
1983 Multiple meanings of 'phoneme' (articulatory, acoustic, perceptual, graphemic) and their confusions. In N. J. Lass (ed.), *Speech and Language: Advances in Basic Research and Practice*. Vol. 9. 285–311. New York: Academic Press.

Warren, R. M., C. J. Obusek, R. M. Farmer and R. P. Warren
1969 Auditory sequence: Confusion of patterns other than speech or music. *Science* 164. 586–587.

Wier, C. C., and D. M. Green
1975 Temporal acuity as a function of frequency difference. *Journal of the Acoustical Society of America* 57. 1512–1515.

Yund, E. W., and R. Efron
1974 Dichoptic and dichotic micropattern discrimination. *Perception and Psychophysics* 15. 383–390.

Part III

Fossil evidence

Appendix B

Broadband repetition pitch: Spectral dominance or pitch averaging?

Richard M. Warren and James A. Bashford, Jr.
Department of Psychology, University of Wisconsin–Milwaukee, P.O. Box 413, Milwaukee, Wisconsin 53201

(Received 13 March 1988; accepted for publication 23 August 1988)

Repetition pitch (RP) produced by mixing noise with its restatement was studied under a variety of delays and filtering conditions. Both normal or cophasic mixtures (RP +) and polarity inverted or antiphasic mixtures (RP −) were used. In keeping with earlier reports, RP + having a delay of t seconds produced a pitch of $1/t$ Hz for all spectral regions examined. Broadband RP − diverged from $1/t$ Hz in keeping with the literature, but the pitches heard under novel filtering conditions indicated that (contrary to some current theories) RP − is a weighted average of the different pitches contributed by different spectral regions. Polarity inversion of an echo introduces additional frequency-dependent delays, and it is suggested that the corresponding RP − at local regions of the basilar membrane reflects a temporal domain analysis based on the sum of these two types of delays.

PACS numbers: 43.66.Hg, 43.66.Ki [NFV]

INTRODUCTION

When a broadband noise is added to itself following a delay of t seconds, a pitch corresponding to $1/t$ Hz can be heard for delays ranging from approximately 0.0005 s (corresponding to 2000 Hz) through 0.02 s (corresponding to 50 Hz) (Fourcin, 1965; Bilsen, 1966; Wilson, 1966). This "repetition pitch" or RP (Bilsen, 1966) has some interesting special characteristics.

When noise is mixed with its echo, the resulting rippled power spectrum has its first spectral peak at $1/t$ Hz and a harmonic succession of peaks at integral multiples of $1/t$ Hz. It seems reasonable to attribute the pitch of rippled noise to spectral cues provided by the loci of stimulation maxima on the basilar membrane. However, we shall see that effects produced by phase shifting suggest that temporal analysis of neural response can play an important role in determining the pitch of rippled noise.

The introduction of a relative phase shift between the delayed and nondelayed components of the rippled stimulus produces a displacement of all spectral peaks by the same absolute value, and results in a change in both pitch value and pitch strength (Fourcin, 1965; Wilson, 1966; Bilsen and Ritsma, 1967/68, 1969/70; Ritsma and Bilsen, 1970; Yost and Hill, 1978; Yost *et al.*, 1978). The phase-shift condition studied most extensively for repetition pitch involves a change of 180 deg (RP −). When the polarity of either component is inverted, the maxima of the spectral ripples are displaced downward in frequency by half the delay reciprocal, so that the peaks of the "antiphasic" RP − stimulus are found at the position of troughs in the corresponding "cophasic" stimulus (RP +). This change in spectral positioning produces a relatively small change in pitch (roughly 10% in most studies). In addition, the pitch is ambiguous, with values of roughly − 10% and + 10% both being heard (Fourcin, 1965; Bilsen, 1966; Wilson, 1966; Bilsen and Ritsma, 1967/68, 1969/70; Yost *et al.*, 1978).

Recent theories of pitch perception applied to both RP + and RP − include a spectral pattern matching model (Bilsen, 1977; Bilsen and Goldstein, 1974) and a filtered autocorrelation model (Yost and Hill, 1978, 1979; Yost *et al.*, 1978; Yost, 1982). These models share a common assumption of "spectral dominance," which considers that "... if pitch information is available along a large part of the basilar membrane the ear uses only the information from a narrow band. This band is positioned at 3 to 5 times *the frequency* value of the pitch" (Ritsma and Bilsen, 1970). That is, each model considers that the pitch of broadband rippled noise is determined by information contained within a narrow frequency band centered in the vicinity of the fourth spectral peak. In this respect, modern theories of RP are similar to pitch theories dealing with the line spectra of complex tones (Goldstein, 1973; Wightman, 1973; Terhardt, 1974, 1979; Srulovicz and Goldstein, 1983).

The pattern matching theory explains the dual pitches of broadband RP − by considering that the spectral information available from the dominant region (neighborhood of the fourth spectral peak) is used for calculation of the fundamental of a cophasic harmonic sequence having peaks at frequencies close to the actual fourth and fifth peaks of the RP − stimulus. This extrapolation has two solutions, resulting in "pseudofundamentals" at approximately $0.9/t$ Hz and $1.1/t$ Hz, in agreement with most empirical findings. Yost's autocorrelation theory considers that temporal information from the fourth spectral mountain is used for an autocorrelational analysis, yielding a two-valued solution equivalent to that resulting from the spectral pseudofundamental calculation.

There are compelling reasons to believe that the region in the vicinity of the fourth and fifth spectral peaks plays an important role in pitch perception (see Plomp, 1976, pp. 114–118). However, it is not at all certain that this region is the exclusive determinant of RP. For example, Wilson (1966) has reported that the dual pitches evoked by broadband RP − vary in the extent of their deviation from $1/t$ Hz as a function of t (with a minimum deviation of approximately 6% at $t = 25$ ms, and a maximum deviation of approximately 20% at $t = 1$ ms). Since the pitches based on calculated pseudofundamentals or autocorrelation peaks for the dominant region of an RP − spectrum deviate from $1/t$ Hz by a constant percentage regardless of t, Wilson's observations would seem to indicate that different spectral regions dominate at different values of t, or that changes in components outside the dominant region can influence broadband RP −.

One of the goals of the present study was to test the validity of spectral dominance theory by measuring RP + and RP − under a variety of filtering conditions. As we shall see, the results obtained indicate that the "dominant" spectral region contributes to, but does not determine, the pitch of broadband RP. The data obtained for RP together with other information in the literature suggest that a temporally based model can provide an explanation for both antiphasic and cophasic repetition pitch.

I. GENERAL METHOD

A. Preparation of stimuli

For the preparation of rippled noise stimuli, white noise produced by a General Radio model 1382 noise generator was bandpass filtered from 50 Hz to 8 kHz (General Radio model 1952 universal filter: 30 dB/oct slopes) and then passed through a custom-modified Eventide model BD955 digital delay line (50-kHz sampling frequency and 10-bit coding) under the control of a Hewlett–Packard model 3325A frequency synthesizer acting as an external clock. The delay line and external clock were adjusted to produce six values of the delay time t, which corresponded to values of $1/t$ Hz ranging in whole-tone steps from 110 to 196 Hz (110, 123, 139, 156, 175, and 196 Hz, respectively). For each value of $1/t$ Hz, the delayed noise was added with unchanged polarity to the nondelayed noise to produce the stimuli for RP +, and with a polarity inversion (performed digitally within the delay line) to produce the stimuli for RP −. The delayed and nondelayed outputs from the delay line were each passed through separate matched Rockland model 852 dual Hi/Lo filters (50 Hz–8 kHz bandpass, with slopes of 48 dB/oct) and, following this identical filtering, were then mixed at equal amplitude by a Gately SPM-6 stereo mixer. These rippled noise stimuli were then passed successively through filters (Rockland model 1042 and Wavetek/Rockland model 751A) to produce the following five spectral ranges for both RP + and RP − (filter slopes for all conditions were 211 dB/oct with cutoff frequencies set at the following positions): broadband (50–8000 Hz); bandpass from the third to the seventh spectral peak; low-pass up to the seventh spectral peak; band-reject between the third and seventh peak; and high-pass from the seventh spectral peak. Since peak frequencies change with delay setting and with polarity inversion, the cutoff frequencies of the filters were adjusted accordingly. In order to avoid edge pitches produced by the steep filter slopes, the rejected spectral components were replaced by uncorrelated white "filler" noise subjected to complementary filtering and having the same spectrum level (dB/Hz) for all conditions (except broadband). All RP stimuli and filler noise bands, including the high- and low-pass components of the band-reject conditions, were recorded on separate tracks of an Ampex MM1200 16-track recorder at 15 ips, and were mixed down during the experiment using a Yamaha model PM-430 audiomixer. The output of the mixer was subjected to a final low-pass filtering at 4 kHz (115 dB/oct slopes) to produce the rippled noise stimuli listed in Tables I and III.

B. Subjects

Five listeners participated in this experiment. Two listeners (CG and JB) had had prior musical training and one (JB) had also had prior experience in repetition pitch matching. All listeners received between 1 and 8 h of training in matching sinusoids to broadband RP + using the method of adjustment. Each delay time was selected randomly from the range of 5–10 ms (20-μs steps), corresponding to pitch values from 200 to 100 Hz, respectively. Listeners began their participation in the formal experiment when their pitch judgments corresponded to $1/t$ Hz within ± 1.5% for each of six successive values of t in three successive blocks of practice trials.

C. Procedure

Preliminary training and formal testing were carried out in an audiometric room, with the rippled noise stimuli presented at 55 dBA SPL through diotically wired TDH-49 headphones. The rippled noise stimuli were matched with sinusoidal tones by the listener, who adjusted the output of a Wavetek 166 function generator using both the main and vernier frequency control dials (calibration marks were concealed from view). The selected frequency match (measured with an accuracy of 0.01 Hz) was recorded by the experimenter, who monitored a Hewlett–Packard model 5316A universal counter/timer, also concealed from the listener's view. During matching, the listener could switch between the rippled noise and the adjustable matching tone at will. The listener could also switch to an on-line white noise presented at the same intensity and having the same bandwidth (50 Hz–4 kHz) as the RP stimulus. This flat spectrum noise served as a neutral buffer, and listeners found it helpful when it was employed prior to the presentation of a new echo delay. No feedback or knowledge of results was provided during the study. Subjects could, at their option, defer matching at any particular value of t, and match at the next scheduled value before returning to the previous stimulus (this option was seldom used more than once per session). Listeners also had the option of canceling a session in progress if they did not wish to continue (this option was exercised eight times out of a total of 293 sessions).

The five filtering conditions were presented in separate segments of the study and in the following order: (1) broadband, (2) high-pass, (3) low-pass, (4) band-reject, and (5) bandpass. Within each filtering condition, listeners completed all matches for RP + before providing matches for RP −. There were five experimental sessions for RP + matching, with listeners producing one match at each of the six delays in each session. Within sessions, the six echo delays were presented in a pseudorandom order, with the restriction that the last delay in one session did not serve as the first delay in the next. The procedure for the RP − conditions was the same, except that some listeners needed more than five sessions. These additional sessions were necessary because of the dual pitches associated with RP −. Matches below $1/t$ Hz were much more frequent than those above (in keeping with reports by Fourcin, 1965; Bilsen, 1966; and Wilson, 1966), and listeners were required to repeat all judgments of an antiphasic condition until they had accumulated five matches below $1/t$ Hz at each value of t.

D. Results

The primary data employed for analysis were the averages of each listener's five adjustments of the matching sinusoidal tone under the various combinations of echo delay, filtering condition, and repetition phase shift (0 or 180 deg). The group results for the matching of RP + under the five filtering conditions are presented in Table I, expressed as the average percent deviation of matches from $1/t$ Hz at each value of t. As shown in Table I, the accuracy of matching was high at all echo delays and under all filtering conditions, with an overall deviation from $1/t$ Hz averaging only 0.52%. The results for individual listeners, averaged across echo delays for each filtering condition, are presented in Table II.

The group results for the matching of the antiphasic RP − stimuli are presented in Table III, and the results for individual listeners are presented in Table IV.

II. DISCUSSION

A. Spectral dominance versus pitch averaging

The data obtained in this study are not in accord with the spectral dominance theory, and an alternative broad spectrum basis for pitch is proposed. Let us first relate our experimental data to predictions based on the spectral dominance theory.

The data gathered for normal or cophasic repetition (RP +) in this study serve mainly as a control for the measurements of the effect of filtering upon antiphasic repetition pitch (RP −). The broadband (unfiltered) pitch judgments for RP + shown in Table I agree closely with the value of $1/t$ Hz (where t is the echo delay in seconds) reported by Bilsen and Ritsma (1969/70) and others. In addition, it was found that RP + judgments approximated $1/t$ Hz for a variety of filtering conditions, including those in which the dominant spectral region for broadband rippled noise was absent. This finding is in keeping with the report that 1/3 octave bands of cophasic rippled noise outside the normally dominant region have values approximating $1/t$ Hz (Bilsen and Ritsma, 1969/70). The fact that other spectral regions produce the same repetition pitch does not conflict with spectral dominance theory since the theory considers that, although the region in the vicinity of the fourth spectral peak is the sole determinant of pitch when present, if absent, then other regions can give rise to repetition pitch. Still, the observation that the same pitch is heard for RP + under various filtering conditions leaves open the possibility that spectral regions outside the range of "dominance" also contribute to the pitch of the broadband stimulus. Filtered antiphasic repetition pitch (RP −) can provide a critical test of this alternative to spectral dominance theory, and the current study was designed to provide such a test.

Our finding that RP − heard broadband differs from that heard for the dominant region when presented alone (bandpass condition) contradicts spectral dominance theory: If the dominant region were the sole determinant of pitch when present, then RP − for the broadband condition and for the dominant region bandpass condition should be the same. Data obtained under other filtering conditions suggest that the pitch heard broadband is based upon a pooling of the different pitches associated with particular spectral regions of the antiphasic spectrum. Thus, while the complementary bandpass and band-reject pitches each differ from the broad-

TABLE I. Mean percent deviation from $1/t$ Hz for matches of pure tones to various spectral ranges of filtered RP +. Results shown are means and standard errors (s.d./$\sqrt{5}$) for the average matches of five listeners.

	1/t Hz						
Filtering of RP +	110	123	139	156	175	196	Grand mean
Broadband (50–4000 Hz)	0.05 (0.08)	0.34 (0.15)	0.53 (0.29)	0.41 (0.13)	0.47 (0.25)	0.68 (0.23)	0.41 (0.09)
Bandpass (3rd–7th peak)	0.55 (0.23)	− 0.03 (0.46)	0.11 (2.24)	0.25 (0.97)	0.39 (0.61)	0.95 (1.74)	0.37 (0.14)
Low-pass (lst–7th peak)	0.32 (0.31)	0.61 (0.23)	1.47 (0.87)	0.60 (0.31)	0.59 (0.13)	2.11 (1.34)	0.95 (0.33)
Band-reject (3rd–7th peak)	0.00 (1.01)	1.02 (1.34)	− 0.70 (0.17)	− 0.34 (0.15)	− 0.77 (0.38)	− 0.42 (0.50)	− 0.20 (0.25)
High-pass (7th peak–4000 Hz)	− 0.92 (0.57)	− 0.28 (0.40)	− 0.41 (0.53)	− 1.72 (0.31)	− 0.04 (0.41)	− 0.63 (0.43)	− 0.67 (0.12)

TABLE II. Mean percent deviation from 1/t Hz for pure-tone matches to RP + under five filtering conditions. Results shown are means and standard errors (s.d./$\sqrt{6}$) for the average matches of individual listeners at six values of t.

	Filtering condition				
Subject	Broadband	Bandpass	Low-pass	Band-reject	High-pass
CG	0.57 (0.14)	0.18 (0.32)	0.82 (0.12)	− 0.54 (0.30)	− 0.90 (0.61)
DG	0.13 (0.15)	0.57 (0.68)	1.23 (0.15)	0.33 (1.17)	− 0.28 (0.14)
JB	0.31 (0.09)	0.36 (0.12)	0.17 (0.19)	− 0.11 (0.15)	0.01 (0.13)
BB	0.67 (0.24)	− 0.41 (0.21)	1.48 (2.45)	− 1.37 (0.21)	− 3.64 (0.76)
MM	0.38 (0.23)	− 1.30 (0.69)	1.05 (0.75)	0.68 (0.70)	1.54 (0.44)
Mean	0.41 (0.09)	0.37 (0.14)	0.95 (0.33)	− 0.20 (0.25)	− 0.67 (0.12)

band pitch, their mean approximates that of the broadband condition, indicating that the effect produced by each alone is averaged when both are present simultaneously. Our data provide an additional example of pitch averaging: The pitch heard for the low-pass condition (which includes the dominant region and contains all peaks up to the seventh) deviates from that heard for the broadband condition by 4 units of standard error, but, when averaged with the pitch associated with the complementary high-pass condition (all peaks down to the seventh), the value once again approximates that obtained for the broadband mixture of the complementary segments.

B. Polarity inversion and local time delays

If a cophasic RP + with a time delay of t seconds is converted to antiphasic RP − by polarity inversion of the delayed sound, then an additional frequency-dependent time delay is introduced. For a 1/3 octave band (approximating a critical bandwidth) with a center frequency of f Hz, an additional delay of plus or minus half of the period of the center frequency is introduced, so that the overall delay is $t \pm 1/2f$ s, and pitch based upon the local repetition time becomes $1/(t \pm 1/2f)$ Hz. Using this simple expression, the calculated values for a decrease in pitch resulting from a polarity inversion at the regions of the fourth, fifth, and sixth peaks are 12.5%, 10.0%, and 8.3%, respectively. The same expression for antiphasic repetition pitch was derived from the major positive peaks in the autocorrelation function (using simplifying assumptions) by Yost *et al.* (1978), and used by them to account for the empirical values for broadband RP − in terms of the pitch at the dominant spectral region.[1]

The empirical pitch values shown in Table III for the various filtering conditions of RP − correspond most closely to the delays calculated for the particular spectral peaks given in parentheses: broadband (sixth); bandpass from third to seventh peaks (fifth); low-pass up to seventh peak (fifth); high-pass from the seventh peak (ninth); band-reject lacking third through seventh peaks (ninth). These values are consistent with the theory that repetition pitch (whether RP + or RP −) is determined by the averaging of local temporally based pitch values.

A time-domain basis for RP is consistent with observations involving long repetition delays. Warren *et al.* (1980) reasoned that, if temporal processing were responsible for repetition pitch, it might be possible to detect repetition for delays extending beyond the limit for pitch even though spectral cues to repetition were unavailable. It was found that delays as long as 0.5 s could be detected and matched

TABLE III. Mean percent deviation from 1/t Hz for lower pitch matches of pure tones to various spectral ranges of filtered RP − . Results shown are means and standard errors (s.d./$\sqrt{5}$) for the average matches of five listeners.

	1/t Hz						
Filtering of RP −	110	123	139	156	175	196	Grand mean
Broadband (50–4000 Hz)	− 7.08 (0.10)	− 7.49 (0.31)	− 8.18 (0.67)	− 9.26 (0.98)	− 9.12 (0.77)	− 8.18 (0.65)	− 8.22 (0.35)
Bandpass (3rd–7th peak)	− 8.62 (0.85)	− 10.07 (0.49)	− 10.28 (1.12)	− 9.66 (1.18)	− 13.98 (2.24)	− 11.91 (5.28)	− 10.75 (0.78)
Low-pass (1st–7th peak)	− 8.53 (0.84)	− 9.62 (0.22)	− 9.79 (0.24)	− 10.00 (0.32)	− 9.55 (0.36)	− 10.30 (0.73)	− 9.63 (0.25)
Band-reject (3rd–7th peak)	− 5.22 (0.38)	− 5.96 (0.36)	− 5.86 (1.07)	− 6.39 (0.52)	− 6.24 (0.69)	− 6.36 (0.72)	− 6.01 (0.18)
High-pass (7th peak–4000 Hz)	− 5.25 (0.40)	− 5.15 (0.20)	− 5.90 (0.38)	− 5.51 (0.51)	− 5.31 (0.51)	− 5.72 (0.57)	− 5.47 (0.12)

TABLE IV. Mean percent deviation from $1/t$ Hz for the lower pitch matches of pure tones to RP — under five filtering conditions. Results shown are means and standard errors (s.d./$\sqrt{6}$) for the average matches of individual listeners at six values of t.

	Filtering condition				
Subject	Broadband	Bandpass	Low-pass	Band-reject	High-pass
CG	− 7.58 (0.45)	− 9.89 (0.22)	− 9.59 (0.31)	− 5.06 (0.19)	− 5.01 (0.12)
DG	− 7.97 (0.33)	− 9.36 (0.44)	− 9.25 (0.21)	− 5.72 (0.12)	− 5.19 (0.31)
JB	− 7.41 (0.14)	− 9.68 (0.16)	− 9.58 (0.10)	− 5.36 (0.11)	− 5.29 (0.19)
BB	− 9.19 (0.93)	− 15.93 (2.72)	− 10.88 (0.44)	− 6.75 (0.87)	− 6.93 (0.30)
MM	− 10.30 (0.54)	− 10.07 (2.01)	− 8.76 (1.26)	− 7.25 (0.91)	− 5.48 (0.79)

accurately. At this duration, neighboring spectral peaks were separated by only 2 Hz, which was much too close to permit resolution on the basilar membrane, and so temporal (autocorrelational?) analysis was responsible for detection of repetition. It was reported by Warren *et al.* that infrapitch repetition was insensitive to polarity inversion, so that antiphasic repetition was indistinguishable from cophasic repetition. This equivalence would be anticipated from a temporal theory since, for the long delays of infrapitch repetition, changes in delay times produced by polarity inversion drop below the just-noticeable difference for all audible spectral regions.

In conclusion, it appears that while the region in the neighborhood of the fourth spectral peak contributes to the repetition pitch heard for broadband stimuli, it does not determine pitch as maintained by the spectral dominance theory. The results reported here for antiphasic repetition pitch, together with other evidence, indicate that the perceived pitch is based upon a pooling of information across critical bands. At each cochlear locus, the effective repetition period responsible for pitch is equal to the sum of the repetition delay and any additional local frequency-dependent delay produced by polarity inversion. The weighted average of these local time delays corresponds to the repetition pitch heard broadband.

ACKNOWLEDGMENTS

The support of a grant from the Air Force Office of Scientific Research is gratefully acknowledged. We thank Bradley S. Brubaker for his valuable contributions.

[1] A temporal explanation for repetition pitch heard for bandpass filtered pulse pairs was proposed in a brief letter published by Bilsen and Ritsma (1967/68). They suggested that repetitive features of the fine structure of the waveform produced at discrete loci on the basilar membrane were responsible for the RPs heard. This temporal theory was subsequently discussed more fully by them (Bilsen and Ritsma, 1969/70), and they explicitly stated, "It is important to note that, in the case of continuous noise with its repetition, Repetition Pitch cannot possibly result from a process of detection of a temporal envelope because this is, essentially, missing . . . " (p.67). However, as pointed out by Yost *et al.* (1978), an autocorrelational analysis of neural patterns of stimulation could be used to determine delay times for iterated continuous noise (and hence RP) at individual loci on the basilar membrane.

Bilsen, F. A. (1966). "Repetition pitch: Monaural interaction of a sound with the repetition of the same, but phase shifted, sound," Acustica **17**, 295–300.

Bilsen, F. A. (1977). "Pitch of noise signals: Evidence for a 'central spectrum,' " J. Acoust. Soc. Am. **61**, 150–161.

Bilsen, F. A., and Goldstein, J. L. (1974). "Pitch of dichotically delayed noise and its possible spectral basis," J. Acoust. Soc. Am. **55**, 292–296.

Bilsen, F. A., and Ritsma, R. J. (1967/68). "Repetition pitch mediated by temporal fine structure at dominant spectral regions," Acustica **19**, 114–115.

Bilsen, F.A., and Ritsma, R. J. (1969/70). "Repetition pitch and its implication for hearing theory," Acustica **22**, 63–73.

Fourcin, A. J. (1965). "The pitch of noise with periodic spectral peaks," in *Reports of the Fifth International Congress on Acoustics*, Liège, Belgium, 1965, 1A, B42, pp. 1–5.

Goldstein, J. L. (1973). "An optimum processor theory for the central formation of the pitch of complex tones," J. Acoust. Soc. Am. **54**, 1496–1516.

Plomp, R. (1976). *Aspects of Tone Sensation* (Academic, London).

Ritsma, R. J., and Bilsen, F. A. (1970). "Spectral regions dominant in the perception of repetition pitch," Acustica **23**, 334–339.

Srulovicz, P., and Goldstein, J. L. (1983). "A central spectrum model: A synthesis of auditory nerve timing and place cues in monaural communication of frequency spectrum," J. Acoust. Soc. Am. **73**, 1266–1276.

Terhardt, E. (1974). "Pitch, consonance, and harmony," J. Acoust. Soc. Am. **55**, 1061–1069.

Terhardt, E. (1979). "Calculating virtual pitch," Hear. Res. **1**, 155–182.

Warren, R. M., Bashford, J. A., Jr., and Wrightson, J. M. (1980). "Infrapitch echo," J. Acoust. Soc. Am. **68**, 1301–1305.

Wightman, F. L. (1973). "The pattern transformation model of pitch," J. Acoust. Soc. Am. **54**, 407–416.

Wilson, J. P. (1966). "Psychoacoustics of obstacle detection using ambient or self-generated noise," in *Animal Sonar Systems*, edited by R. G. Busnel (Louis-Jean, Gap, Hautes-Alpes, France), pp. 89–114.

Yost, W. A. (1982). "The dominance region and ripple noise pitch: A test of the peripheral weighting model," J. Acoust. Soc. Am. **72**, 416–425.

Yost, W. A., and Hill, R. (1978). "Strength of pitches associated with ripple noise," J. Acoust. Soc. Am. **64**, 485–492.

Yost, W. A., and Hill, R. (1979). "Models of the pitch and pitch strength of ripple noise," J. Acoust. Soc. Am. **66**, 400–410.

Yost, W. A., Hill, R., and Perez-Falcon, T. (1978). "Pitch and pitch discrimination of broadband signals with rippled power spectra," J. Acoust. Soc. Am. **63**, 1166–1173.

Appendix C

Illusory continuity of tonal and infratonal periodic sounds

Richard M. Warren, John M. Wrightson,[a] and Julian Puretz
Department of Psychology, University of Wisconsin–Milwaukee, P.O. Box 413, Milwaukee, Wisconsin 53201

(Received 20 August 1987; accepted for publication 22 June 1988)

Temporal induction can restore masked or obliterated portions of signals so that tones may seem continuous when alternated with sounds having appropriate spectral composition and intensity. The upper intensity limits for the induction of tones (pulsation thresholds) are related to masking functions and have been used to define the characteristics of frequency domain (place) analysis of tones. The present study has found that induction also occurs for infratonal periodic sounds that require a time domain analysis for perception of acoustic repetition. Limits for temporal induction were determined for iterated frozen noise segments from 10–2000 Hz alternated with a louder on-line noise. Masked thresholds were also obtained for the pulsed signals presented along with continuous noise, and it was found that the relation between induction limits and masking changed with frequency. The results obtained for induction and masking are discussed in terms of general principles governing restoration of obliterated sounds.

PACS numbers: 43.66.Mk, 43.66.Dc [NFV]

INTRODUCTION

When portions of signals are replaced by louder sounds, listeners may believe they hear the missing fragments. The limiting conditions for these perceptual restorations have been studied in recent years as a source of information concerning auditory mechanisms.

Miller and Licklider (1950) appear to have published the first report of illusory continuity of signals interrupted by noise. They found that when two sounds differing in intensity and quality are alternated in a regular fashion, the fainter sound may seem to remain on continuously. Thurlow's (1957) rediscovery of this effect led to a number of subsequent studies (Thurlow and Elfner, 1959; Thurlow and Marten, 1962; Elfner and Caskey, 1965; Elfner and Homick, 1966, 1967; Elfner, 1969, 1971).

Houtgast (1972) and Warren *et al.* (1972) independently proposed rules considering the apparent continuity of the fainter of two alternating sounds as the inverse of masking. Houtgast's rule describing what he called the *pulsation threshold* for tones was: "When a tone and a stimulus S are alternated (alternation cycle about 4 Hz), the tone is perceived as being continuous when the transition from S to tone causes no (perceptible) increase of nervous activity in any frequency region. The pulsation threshold, thus, is the highest level of the tone at which this condition still holds." This rule has been used to infer the characteristics of spectral filtering at the basilar membrane, with experimental findings being interpreted in terms of both topographical excitation patterns produced by tones and lateral suppression at loci contiguous to the stimulated regions (see, for example, Houtgast, 1974; Aldrich and Barry, 1980; Shannon and Houtgast, 1986).

The rule for *temporal induction* described by Warren *et al.* (1972) also requires an overlap of peripheral neural excitation but is somewhat different in scope, stating that: "If there is contextual evidence that a sound may be present at a given time, and if the peripheral units stimulated by a louder sound include those which would be stimulated by the anticipated fainter sound, then the fainter sound may be heard as present." This principle applies not only to the perceptual restoration of fragments of steady-state tones but also to the restoration of several types of time-varying signals which include speech, tonal glides, and melodic tonal sequences (see Warren, 1984). The present investigation extends the study of temporal induction to complex sounds having repetition frequencies below the limit of pitch. These long-period sounds are perceived as possessing a repetitive temporal texture or time-varying pattern.

In their investigation of such sounds, Guttman and Julesz (1963) used iterated segments of Gaussian noise (repeated "frozen noises" or RFNs). They described the perceptual quality as a repetitive "whooshing" from 1 Hz (the approximate lower limit of periodicity detection of RFNs) to 4 Hz and as "motorboating" from 4–19 Hz. At 20 Hz and above, RFNs are considered to be complex tones possessing pitch and were not investigated by Guttman and Julesz. Warren and Bashford (1981) examined both tonal and infratonal RFNs, and reported that a noisy pitch with a hiss-like quality was heard from 20 Hz up to about 100 Hz, with RFNs of higher frequencies appearing to be completely tonal with no hint of a noisy quality. They noted similarities in the rules governing perception of pitch and infrapitch, and suggested that some of the mechanisms for the detection of acoustic repetition may operate on both sides of the tonal/infratonal boundary. It was suggested that RFNs could serve as useful model stimuli for studying the continuum of detectable acoustic iterance, with observations in the infratonal and the tonal ranges each enhancing understanding of the other (for further discussion, see Warren, 1982, pp. 78–90).

The present study was designed to compare the upper limits of temporal induction for complex tones and for infra-

[a] Present address: Joiner–Rose Group, 4125 Centurion Way, Dallas, TX 75244.

tonal periodic sounds. Since experiments have demonstrated a close relation between masking and illusory continuity (Houtgast, 1972; Warren *et al.*, 1972), in the present study both induction limits and various types of masking limits were measured on both sides of the pitch boundary.

I. METHOD

A. Subjects

Four listeners participated in the study. Each was familiar with psychoacoustic experimentation and had served as a subject in other studies of auditory perception.

B. Stimuli

The periodic stimuli consisted of repeated frozen noises (RFNs). The output voltage from a Gaussian noise generator was sampled every 20 μs and coded in 12-bit form by a digital delay line built to our specifications by the Physical Data Company. The delay was adjusted to correspond to the desired period, and then by closing a "recycle" switch, input to the delay line was rejected, and the signal looped or repeated indefinitely in digital form. Appropriate filtering removed the spectral artifacts associated with digital processing. The RFNs had periods of 100, 50, 20, 10, 5, 2, 1, and 0.5 ms which corresponded to frequencies of 10, 20, 50, 100, 200, 500, 1000, and 2000 Hz, respectively. Two modes of stimulus presentation were used: For one mode, the periodic signal was alternated with on-line noise (the signal and the noise were each on for 300 ms); for the other mode, the noise was on continuously, and the signal was pulsed (the superimposed signal was on for 300 ms and off for 300 ms). The on-line noise was always delivered at 80 dB SPL, and the intensity of the periodic signal was adjusted to a particular criterion level by the listener. Timing was controlled by a preset counter driven by a Rockland 5100 frequency synthesizer, and electronic switches used for alternating the stimuli were set for a linear rise–fall time of 25 ms. For repetition frequencies of 50 Hz and above, both the noise and the RFNs were high-pass filtered at the lowest frequency of the periodic signal (the spectral fundamental) and low-pass filtered at 8000 Hz, using filters with slopes of 48 dB/oct. At repetition frequencies below the 50-Hz response limit of both the delay line and the headphones, high-pass filtering was maintained at 50 Hz and the low-pass filtering was again 8000 Hz for both the signal and the on-line noise. The intensity of the RFNs was increased or decreased as desired by the listener by turning the unseen dial of an attenuator having 1-dB steps. Stimuli were presented diotically through matched TDH-49 headphones having a flat response (± 1 dB) from 50–8000 Hz while listeners were seated in an audiometric room having an ambient SPL of 25 dBA.

C. Procedure

Listeners were presented with each of the eight signal frequencies once during an experimental session. At each frequency, they were instructed to make the five different types of judgments described in detail below. For the first three judgments, the RFNs were alternated with 80-dB SPL on-line noise, with each on for 300 ms before switching. For the last two judgments, the 80-dB noise was continuous, and the mixed (added) periodic sound was alternately on for 300 ms and off for 300 ms.

Five types of judgments were made in the following order.

(1) Continuity/discontinuity transition (upper limit of temporal induction): The intensity of the RFN alternated with noise was adjusted to the lowest level at which it seemed discontinuous or pulsant (just below this limit, listeners reported hearing continuous iterance that was either pitch or motorboating).

(2) Threshold for detection of signal presence when alternated with noise: The RFN was adjusted to the lowest level at which its presence could be detected (i.e., it was noticeably different from silence).

(3) Threshold for detection of signal repetition when alternated with noise: The RFN was adjusted to the lowest level at which iteration (either pitch or infrapitch repetition) could be heard.

(4) Threshold for detection of signal presence when superimposed upon noise: The RFN was adjusted to the lowest level at which its presence could be detected as an intermittent addition to the continuous noise.

(5) Threshold for detection of signal repetition when superimposed upon noise: The RFN was adjusted to the lowest level at which iteration (either pitch or infrapitch) could be heard for the intermittent addition to the continuous noise.

There were six experimental sessions. Each session was split into two parts separated by a 5-min rest period. During part A, listeners were presented with four of the eight RFN frequencies (10, 50, 200, and 1000 Hz) presented in a randomly determined order. The five types of judgments described above were made successively in the order listed for each of the frequencies. Part B was the same as part A, except that the remaining four repetition stimulus frequencies were employed (20, 100, 500, and 2000 Hz). In the first, third, and fifth sessions, part A was presented first, followed by part B. In the second, fourth, and sixth sessions, this order was reversed. By the end of the study, each listener had made six judgments for each of the five types of thresholds with each of the eight iterated noise segment frequencies. It should be noted that each frozen waveform was used for only one session and one listener.

II. RESULTS

The experimental data obtained are summarized in Table I. It can be seen that induction was greatest (the continuity/discontinuity boundaries were at the highest amplitudes) at infratonal and low tonal repetition frequencies. A one-way analysis of variance with repeated measures yielded a significant effect of frequency [$F(7,21) = 20.53$, $p < 0.001$], and subsequent Newman–Keuls tests indicated that continuity/discontinuity thresholds were higher ($p < 0.05$ or better) at repetition frequencies from 10–100 Hz than at repetition frequencies from 200–2000 Hz.

Table I also shows that, when the RFN was alternated with noise, the threshold for detecting signal repetition was

TABLE I. Different types of thresholds for signals consisting of 300-ms bursts of iterated noise segments when alternated with, or superimposed upon, on-line noise at 80 dB SPL. Means and standard error (SE) of means are in dB SPL and represent 24 judgments (six from each of four subjects). For further details, see text.

	Repetition frequency of iterated noise segment (Hz)							
	10	20	50	100	200	500	1000	2000
	Continuity/discontinuity transition (alternation) (SPL)							
Mean	73.38	71.79	68.08	66.00	59.79	57.58	53.67	53.63
SE	0.49	0.93	0.85	0.89	1.65	1.88	1.49	1.78
	Detection of signal presence (alternation) (SPL)							
Mean	31.21	30.50	29.50	30.25	28.46	28.88	28.00	30.46
SE	0.80	0.90	0.81	0.70	0.75	0.85	0.73	0.98
	Detection of signal pitch or motorboating (alternation) (SPL)							
Mean	36.58	34.88	32.29	31.71	29.29	29.71	28.13	31.13
SE	1.34	1.28	1.08	0.82	0.79	0.83	0.65	0.95
	Detection of signal presence (simultaneous) (SPL)							
Mean	70.54	70.58	69.67	69.46	68.21	66.63	65.08	64.75
SE	0.44	0.55	0.51	0.56	0.47	0.55	0.81	0.57
	Detection of signal pitch or motorboating (simultaneous) (SPL)							
Mean	75.58	75.29	72.21	70.92	68.63	66.79	65.50	65.54
SE	0.45	0.57	0.88	0.56	0.60	0.66	0.82	0.66

several dB above the threshold for detecting signal presence for the infratonal and low tonal frequencies. Planned orthogonal comparisons (Kirk, 1968, pp. 73–76) indicated that the two thresholds differed reliably ($p < 0.01$) at repetition frequencies of 10, 20, and 50 Hz. As the intensity level was raised for these low frequencies, the repeated frozen noise (RFN) was heard first only as a faint continuous hiss without detectable iteration: An appreciable increase in amplitude (3–5 dB) above the absolute detection threshold was required before effects attributable to repetition could be heard. The boundary between pitch which seems completely homogeneous and tonal and pitch with a noisy or hisslike quality occurs at about 100 Hz (Warren and Bashford, 1981). As shown in Table I, the pitch corresponding to these purely tonal RFNs was detected at signal intensities approximating the absolute detection threshold.

Although simultaneous masking was absent when the periodic sounds were alternated with on-line noise, the possibility of forward and backward masking produced by noise bursts preceding and following the signal needs to be considered. The 300-ms duration of interruptions used for both induction and threshold measurements would be expected to produce only a slight, if any, increase in thresholds in the present study (for a discussion of the limits of forward and backward masking and their interactions, see Elliot, 1971; Wilson and Carhart, 1971). Nevertheless, in order to compensate for any residual masking of this type, the threshold for detection of repetition when the signal was alternated with noise was subtracted from the amplitude corresponding to the continuity/discontinuity transition to obtain the sensation level (SL) at the upper limit of auditory induction for each stimulus frequency. These values were used to construct Fig. 1 showing the existence regions for temporal induction (illusory continuity of acoustic repetition) and for pulsation (perception of discontinuity). Listeners' SLs for induction were subjected to an analysis of variance that yielded a significant effect of frequency [$F(7,21) = 7.08$, $p < 0.001$]. Subsequent Newman–Keuls tests indicated that the existence region for iterance was diminished ($p < 0.05$) for repetition frequencies of 1–2 kHz.

The correspondence between the continuity/discontinuity transition and the threshold for detection of repetition under conditions of simultaneous masking is shown in Fig. 2. The data for these two thresholds were compared in a two-factor analysis of variance that yielded significant main effects of threshold type [$F(1,3) = 16.40, p < 0.05$] and repetition frequency [$F(7,21) = 38.53$, $p < 0.0001$] and a significant interaction [$F(7,21) = 4.20$, $p < 0.005$]. Subsequent Newman–Keuls tests indicated that thresholds for repetition detection under simultaneous masking were higher ($p < 0.05$ or better) than the continuity/discontinuity transition at corresponding repetition frequencies of 200 Hz and above, but these two measures did not differ reliably at repetition frequencies of 100 Hz and below. Thus noise was a relatively poor inducer of continuity for purely tonal sounds, in keeping with the data reported by Warren *et al.* (1972) for 300-ms sinusoidal tones alternated with 300-ms noises.

III. DISCUSSION

Three types of temporal induction have been described: homophonic, contextual catenation, and heterophonic

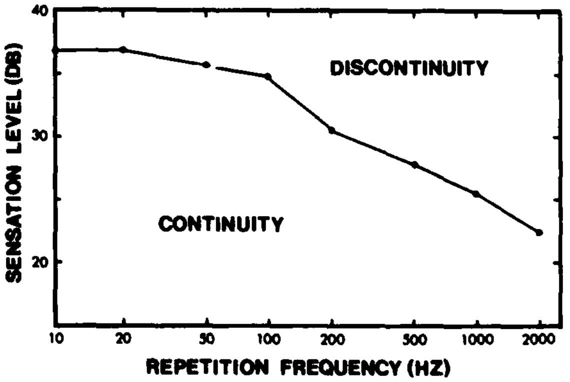

FIG. 1. The discontinuity/continuity boundary in sensation level (dB above repetition detection threshold) for frozen noise segments with different repetition frequencies when alternated each 300 ms with 80-dB SPL on-line noise. For further description, see the text.

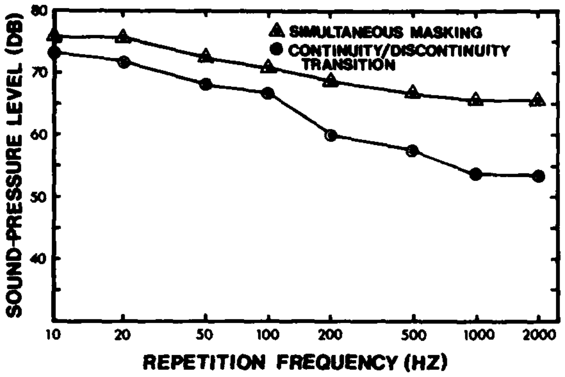

FIG. 2. Comparison of the continuity/discontinuity boundary (upper limit of temporal induction) with the signal threshold under simultaneous masking. The continuity/discontinuity threshold is for the iterated signal when alternated each 300 ms with an 80-dB broadband noise, and the masked threshold is for the detection of signal repetition when added intermittently (on 300 ms and off 300 ms) to a continuous 80-dB broadband noise. All values are in dB SPL. For further details, see the text.

(Warren, 1984). Homophonic induction is the simplest, and its characteristics can facilitate understanding of the others. It occurs when two intensity levels of otherwise identical sounds are alternated, and consists of the apparent continuity of the fainter level. The sounds producing homophonic induction can be periodic (such as two levels of a sinusoidal tone) or nonperiodic (such as two levels of a noise)—in each case, induction of the fainter occurs at all audible differences for all audible levels. It seems that the segments of the weaker sound occurring before and after each segment of the louder sound cause it (the louder sound) to be factored into two portions. One of these portions corresponds to the level of the fainter sound and provides the bridging continuity, while the residue (the original louder level minus the fainter level) appears as a pulsed addition to the continuous sound. A simple demonstration of this subtractive factoring is provided by the observation that, when 80- and 82-dB levels of the same noise are alternated and listeners perceive the 80-dB level as continuous, they paradoxically hear the 82-dB level as a pulsed *fainter* sound (Warren, 1982, p. 141). While the subtractive nature of induction is not as obvious when the inducer and inducee are qualitatively different, there is evidence that the two other types of temporal induction also involve a subtractive processing that may reverse the effects of masking (Warren, 1984).

Contextual catenation occurs when a time-varying signal such as speech is interrupted by a louder extraneous sound. When the peripheral neural overlap requirements (discussed earlier) are met, the contextual information provided by the intact segments can lead to perceptual synthesis of fragments differing from the preceding and following portions of the signal. Listeners hear the signal as uninterrupted and cannot distinguish the restored segments from those physically present. In addition to phonemic restorations (Warren, 1970; Bashford and Warren, 1987), contextual catenation can restore missing notes of a melody played on the piano (Sasaki, 1980) and can synthesize obliterated segments of tonal frequency glides (Dannenbring, 1976; Ciocca and Bregman, 1987).

Heterophonic continuity refers to the apparent lack of interruption of a particular sound when replaced by a qualitatively different louder sound that meets the specifications of the peripheral overlap rule. Tones are often employed as the fainter sound, but other periodic sounds can be employed. The iteration of nonsinusoidal waveforms can be detected at infratonal frequencies (below 20 Hz), and the present study has examined the induction of tonal and infratonal repeated frozen noises over a range extending from 10–2000 Hz.

Let us compare the auditory mechanisms employed for the detection of repetition in the tonal and infratonal ranges, and their relevance to the observations made in the present study. In the infratonal range, perception of frozen noise repetition is based upon the iteration of neural response patterns. This temporal information is available at all loci on the basilar membrane, for, when a ⅓-oct bandpass filter (approximating a critical band) is swept through the audible range, then an infrapitch repetition (attributable to the interaction of unresolved harmonics within a critical band) can be heard at all center frequencies of the filter (Warren and Bashford, 1981). As the RFN frequency is raised into the tonal range, then individual lower harmonics can be resolved along the basilar membrane (see Plomp, 1964, for the limits of spectral resolution), and two additional neural correlates of RFN repetition appear along with the iterated neural patterns corresponding to the unresolved higher harmonics (Warren, 1982, pp. 82–85). The resolved harmonics can provide spectral information concerning RFN repetition frequency through the positioning of stimulation maxima on the basilar membrane, and may also provide temporal information based upon the phase locking of nerve fiber responses (for a discussion of place cues and phase-locked cues to the pitch of complex tones, see de Boer, 1976, and Evans, 1978). It appears that once the peripheral overlap rule is satisfied, then gaps in the cues to repetition do not interfere with the apparent continuity of repeated frozen noises: The perceptual synthesis of RFNs restores all of the qualitative attributes of repetition.

As can be seen in Fig. 1, illusory continuity of iterance occurred at higher sensation levels for repetition frequencies from 10–100 Hz than for frequencies from 200–2000 Hz. This change in induction limits was both monotonic and gradual, and not related in any direct fashion to the pitch/infrapitch transition at 20 Hz. The lower pulsation thresholds at higher frequencies may be attributable to the increase in spacing between harmonic components. This greater frequency separation enhances spectral resolution and concentrates stimulation at those neurons with characteristic frequencies close to the resolved harmonics. The concentration of spectral power at discrete loci would necessitate a drop in level of a tonal RFN in order for the 80-dB noise bursts to satisfy induction's peripheral overlap rule.

Figure 2 shows that the transition from induction to pulsation of an RFN remained close to the simultaneous masking threshold for infratonal and low tonal frequencies. The transition of RFNs from noisy tones to smooth, homogeneous tones occurs at about 100 Hz (Warren and Bashford, 1981), and it can be seen that the pulsation thresholds diverged from masked thresholds above that repetition frequency. A similar separation of masking and pulsation limits for pure tones induced by noise has been reported by Warren *et al.* (1972). They alternated 300-ms bursts of tones and noises of various spectral compositions and found that induction limits were 10 dB or more below the simultaneous masking limits.

Why do pulsation thresholds diverge from simultaneous masking thresholds for tonal induction by noise? One possible explanation starts by considering that pulsation thresholds represent the lower limit for detecting signal *absence* in noise. Noises are characterized by rapid changes in amplitude that produce rapidly fluctuating levels of neuronal stimulation. If, when the noise is present, neurons that had been responding previously to a steady tone exhibit momentary dips below the activity levels corresponding to the tone, then absence of that tone is signaled, and induction is blocked. Hence, pulsation thresholds of the tones were not determined by the average SPL of the noise (as appears to be the case for the simultaneous masking of tone by noise), but rather by transitory minima in the noise power spectrum. When the inducee was a periodic sound that itself had a noiselike quality (RFNs up to 100 Hz), it appears that the brief dips in the amplitude of the on-line noise inducer did not block induction, and that the average sound-pressure levels of the two fluctuating sounds determined both induction limits and simultaneous masking limits.

ACKNOWLEDGMENTS

The support of the Air Force Office of Scientific Research is gratefully acknowledged. We thank James A. Bashford, Jr. and Bradley S. Brubaker for their valuable contributions.

Aldrich, W. M., and Barry, S. J. (1980). "Critical bandwidths measured by the pulsation-threshold technique," J. Aud. Res. 20, 137–141.

Bashford, J. A., Jr., and Warren, R. M. (1987). "Multiple phonemic restorations follow the rules of auditory induction," Percept. Psychophys. 42, 114–121.

Ciocca, V., and Bregman, A. S. (1987). "Perceived continuity of gliding and steady-state tones through interrupting noise," Percept. Psychophys. 42, 476–484.

Dannenbring, G. L. (1976). "Perceived auditory continuity with alternately rising and falling frequency transitions," Can. J. Psychol. 30, 99–114.

de Boer, E. (1976). "On the 'residue' and auditory pitch perception," in *Handbook of Sensory Physiology, Vol. V, Auditory System, Part 3: Clinical and Special Topics*, edited by W. D. Keidel and W. D. Neff (Springer, Berlin), pp. 479–583.

Elfner, L. F. (1969). "Continuity in alternately sounded tone and noise signals in a free field," J. Acoust. Soc. Am. 46, 914–917.

Elfner, L. F. (1971). "Continuity in alternately sounded tonal signals in a free field," J. Acoust. Soc. Am. 49, 447–449.

Elfner, L. F., and Caskey, W. E. (1965). "Continuity effects with alternately sounded noise and tone signals as a function of manner of presentation," J. Acoust. Soc. Am. 38, 543–547.

Elfner, L. F., and Homick, J. L. (1966). "Some factors affecting the perception of continuity in alternately sounded tone and noise signals," J. Acoust. Soc. Am. 40, 27–31.

Elfner, L. F., and Homick, J. L. (1967). "Auditory continuity effects as a function of the duration and temporal location of the interpolated signal," J. Acoust. Soc. Am. 42, 576–579.

Elliot, L. L. (1971). "Backward and forward masking," Audiology 10, 65–76.

Evans, E. F. (1978). "Place and time coding of frequency in the peripheral auditory system: Some physiological pros and cons," Audiology 17, 369–420.

Guttman, N., and Julesz, B. (1963). "Lower limit of auditory periodicity analysis," J. Acoust. Soc. Am. 35, 610.

Houtgast, T. (1972). "Psychophysical evidence for lateral inhibition in hearing," J. Acoust. Soc. Am. 51, 1885–1894.

Houtgast, T. (1974). "Masking patterns and lateral inhibition," in *Facts and Models in Hearing*, edited by E. Zwicker and E. Terhardt (Springer, Berlin), pp. 258–265.

Kirk, R. E. (1968). *Experimental Design: Procedures for the Behavioral Sciences* (Brooks/Cole, Monterey, CA).

Miller, G. A., and Licklider, J. C. R. (1950). "The intelligibility of interrupted speech," J. Acoust. Soc. Am. 22, 167–173.

Plomp, R. (1964). "The ear as a frequency analyzer," J. Acoust. Soc. Am. 36, 1628–1636.

Sasaki, T. (1980). "Sound restoration and temporal localization of noise in speech and music sounds," Tohoku Psychol. Folia 39, 79–88.

Shannon, R. V., and Houtgast, T. (1986). "Growth of pulsation threshold of a suppressed tone as a function of its level," Hear. Res. 21, 251–255.

Thurlow, W. R. (1957). "An auditory figure-ground effect," Am. J. Psychol. 70, 653–654.

Thurlow, W. R., and Elfner, L. F. (1959). "Continuity effects with alternately sounding tones," J. Acoust. Soc. Am. 31, 1337–1339.

Thurlow, W. R., and Marten, A. E. (1962). "Perception of steady and intermittent sound with alternating noise-burst stimuli," J. Acoust. Soc. Am. 34, 1853–1858.

Warren, R. M. (1970). "Perceptual restoration of missing speech sounds," Science 167, 392–393.

Warren, R. M. (1982). *Auditory Perception: A New Synthesis* (Pergamon, Elmsford, NY).

Warren, R. M. (1984). "Perceptual restoration of obliterated sounds," Psychol. Bull. 96, 371–383.

Warren, R. M., and Bashford, J. A., Jr. (1981). "Perception of acoustic iterance: Pitch and infrapitch," Percept. Psychophys. 29, 395–402.

Warren, R. M., Obusek, C. J., and Ackroff, J. M. (1972). "Auditory induction: Perceptual synthesis of absent sounds," Science 176, 1149–1151.

Wilson, R. H., and Carhart, R. (1971) "Forward and backward masking: Interactions and additivity," J. Acoust. Soc. Am. 49, 1254–1263.

Appendix P

Illusory continuity of interrupted speech: Speech rate determines durational limits

James A. Bashford, Jr., Mark D. Meyers, Bradley S. Brubaker, and Richard M. Warren
Department of Psychology, University of Wisconsin—Milwaukee, Milwaukee, Wisconsin 53201

(Received 17 March 1987; accepted for publication 2 August 1988)

Deleted segments of speech can be restored perceptually if they are replaced by a louder noise. An earlier study of this "phonemic restoration effect" found that, when recorded discourse was interrupted periodically by noise, the durational limit for illusory continuity corresponded to the average word duration. The present study employed a different passage of discourse recorded by a different speaker. Durational limits for apparent continuity of discourse interrupted by noise were measured at the normal (original) playback speed, as well as at rates that were 15% greater and 15% less. At the normal playback rate, once again the limit of continuity approximated the average word duration—but of especial interest was the finding that changes in playback rate produced proportional changes in continuity limits. These results, together with other evidence, suggest that phonemic restorations represent a special linguistic application of a general auditory mechanism (auditory induction) producing appropriate syntheses of obliterated sounds, and that for discourse the limits of illusory continuity correspond to a fixed amount of verbal information, and not a fixed temporal value.

PACS numbers: 43.71.Es, 43.70.Fq

AFOSR·TR· 89-1677

INTRODUCTION

When portions of an acoustic signal (either verbal or nonverbal) are removed and replaced by a louder extraneous sound, the fragments are restored perceptually if certain conditions are met. This "continuity effect," also known as "auditory induction," requires that deleted portions of the signal be replaced by a potential masker (Miller and Licklider, 1950; Bashford and Warren, 1987; Houtgast, 1972; Verschuure, 1978; Warren *et al.*, 1972). Under these conditions, illusory continuity may persist through noise-filled gaps lasting several hundreds of ms or more and may involve either the simple continuation of steady-state signals such as tones, or the reconstruction of portions of time-varying signals such as the "phonemic restoration" of interrupted speech (Warren and Obusek, 1971). For a review of the literature and theory encompassing both verbal and nonverbal auditory induction, see Warren (1984).

Bashford and Warren (1987) conducted two experiments in which listeners were presented with speech interrupted by louder noise and were required to adjust the duration of periodic gaps to their thresholds for detecting speech deletion (the upper limit of phonemic restoration). In one experiment, recorded discourse (a passage from an article in a popular magazine) was bandpass filtered (remaining intelligible) and then interrupted by silence or by a bandpass-filtered noise. When the discourse was interrupted by silence, the threshold for detection of gaps averaged about 75 ms. However, when the speech band was interrupted by a louder band of noise having the same center frequency (1.5 kHz) and a slightly greater bandwidth, the threshold gap duration increased dramatically to 304 ms, a value almost exactly equal to the average word duration in the passage (306 ms, discounting pause time). Further, the differential efficacy in producing phonemic restorations for other noise bands having different center frequencies paralleled their potential for masking the speech signal.

In a second experiment, Bashford and Warren presented listeners with broadband speech of three types: (1) an unfiltered version of the discourse passage used in the first experiment; (2) the same discourse passage presented at the same word rate but read with the order of words reversed (intonation and phrasing approximated that of discourse); and (3) lists of isolated monosyllabic words. Threshold gap durations were equivalent (about 50 ms) for each of the three types of speech when the signals were interrupted by silence. However, there was a differential increase in threshold durations when gaps in the stimuli were filled with a broadband noise matching the spectra of the speech signals and having a greater amplitude. Threshold gap durations for isolated monosyllables and for the discourse passage read with backward word order both increased by about 100 ms when gaps in these stimuli were filled with noise. In marked contrast, when noise was added to gaps in the normal reading of the discourse passage, continuity threshold durations increased about 250 ms above the value found for silence, and, once again (as in the first experiment using filtered discourse with spectrally matched interpolated noise), the threshold gap duration was almost exactly equal to the average word duration.

The manipulation of linguistic context in the study by Bashford and Warren produced substantial variations in thresholds for discontinuity. These findings led the investigators to suggest that the durational limits for induction may provide a sensitive measure of the effect of context upon the size of linguistic chunks employed in the perceptual organization of speech. Of especial interest to the present study was the observation that the upper limit for induction with discourse was equivalent to the average word duration. However, even though it appears clear that context does influence the size of speech fragments subject to restoration, it is possible that the close correspondence found between the upper limit of auditory induction and the average duration of words was fortuitous, and the consequence of a general durational limit for illusory continuity of discourse. The present study was designed to determine whether the upper limit for continuity of discourse has fixed temporal constraints or varies with the rate of delivery (and hence the duration of linguistic segments). A different discourse passage was recorded by a different speaker and played back at three rates. In one condition, the average word duration was approximately the same as in the earlier study. In the remaining two conditions, durations of components within the passage were expanded or compressed by 15% to determine whether the durational limit of induction would covary with signal rate.

I. METHOD

A. Subjects

The 40 subjects (18 men and 22 women) were enrolled in the Introductory Psychology course at the University of Wisconsin—Milwaukee and were either given course credit or paid for their participation in the study. They were selected from a larger pool of listeners on the basis of an audiometric screening task described in the procedure section.

B. Stimuli

The speech stimulus was a passage from the U.S. Constitution. The reading was produced in a sound-attenuating chamber (IAC series 400 A) by a male speaker having a general American dialect. The passage was initially recorded using a Sony model F-98 cardioid microphone and a Sony model TC-40 cassette recorder that was equipped with an automatic gain control. This initial recording was then bandpass filtered from 200–5000 Hz with slopes of 48 dB/oct (Rockland model 852 filter) and then rerecorded at three different tape speeds on separate tracks of an Ampex 440-C 8-track recorder equipped with a continuously adjustable speed control. One version of the passage was recorded at the same tape speed used for playback (7½ ips). The two remaining versions of the passage were recorded with tape speed altered so that, upon playback at 7½ ips, one version was heard at a rate 15% greater and the other at a rate 15% less than that of the original recording. The speech sounded normal at each of the three rates, but differences in playback speed produced differences in the spectra of the stimuli. In order to provide spectrally matched noise for each stimulus, pink noise (that is, noise with equal power per octave which approximates the long-term average spectrum of speech) was subjected to the identical bandpass filtering employed for the original recording of the speech (200–5000 Hz) and then recorded on separate tracks of the multitrack recorder at three different speeds (7½ ips, 7½ ips + 15%, and 7½ ips − 15%).

When presented at its original rate, the discourse passage lasted 39 min and had an overall word rate of 185 wpm. The percentage of pause time in the passage was 8.6% as determined through measurements of amplitude-level tracings (Brüel & Kjaer model 2305 graphic level recorder with pen speed of 4 mm/s and paper speed of 1 mm/s). The average word duration (with pause time discounted) was calculated to be 296 ms for the normal version of the passage, 252 ms for the accelerated version of the passage, and 340 ms for the decelerated version. Amplitude fluctuations were also determined graphically for each speech recording (pen speed 4 mm/s, paper speed 0.3 mm/s) and were found to be equivalent for corresponding portions of the passage at each signal rate.

C. Apparatus

The six signals recorded on the multitrack recorder (three of which were speech and three noise as described above) were fed to separate subchannels of a Yamaha PM-430 8-channel mixer. The desired speech signal and its matching noise band were passed from separate master outputs of the mixer to individual electronic switches (Grason–Stadler model 1287-B). The two switches were set for 10-ms rise/fall and were triggered alternately, with a 50% duty cycle, by pulses from a Grason–Stadler model 1219 sequence counter. The sequence counter was driven by a Grason–Stadler model 1270 level zone detector that produced logic pulses at a rate determined by the square wave input from a Wavetek model 135 function generator. During the experiment, this generator, with its dial hidden from view, was adjusted by listeners to vary the rate (duration) of speech interruption. The control knob produced a linear change in interruption rate with turning angle over a range of 0.40–25 interruptions per second (ips). The corresponding durations of speech off-time and on-time during each cycle of interruption, ranged from 1250–20 ms, as measured with an accuracy of 0.01 ms by a Hewlett–Packard 5321-A frequency counter. The alternately gated signals from the two electronic switches were combined with a Grason–Stadler model 1292 passive mixer, passed through an impedance-matching transformer (Grason–Stadler model E10589A), and finally transduced diotically through a matched pair of Telephonics TDH-49 headphones mounted in MX 41/AR cushions. The stimuli were presented at an average (C-scale) amplitude of 62 dB for speech and 72 dB for noise, as measured with a Brüel & Kjaer model 2204 sound level meter equipped with a 6-cc earphone coupler and operating in slow response mode.

D. Audiometric screening

At least 1 day prior to participation in the formal experiment, listeners were screened individually in an IAC single-walled sound-attenuating chamber. A Békésy-type tracking procedure was used with a diotically presented sinusoidal tone changing from 500 Hz to 8 kHz in alternately ascending and descending frequency sweeps of 1 oct/min. Subjects tracked their thresholds by pressing and releasing a remote control switch for the audiometer (Grason–Stadler model E-800), which produced a decrease or increase in tonal intensity at a rate of 2.5 dB/s. Listeners having threshold tracings deviating by more than 22.5 dB from normal at any frequency were not included in the formal experiment. Under these criteria for screening (which were chosen to exclude not only listeners with hearing impairments, but also those who failed to follow the standard audiometric instructions for threshold tracking), approximately 50% of the listeners qualified for further participation in the study.

E. Procedure

Subjects were told that they would be listening to passages from the U.S. Constitution and that their task would be to adjust a dial to the point where interruptions of the voice became clearly detectable. After the experimenter presented them with samples of discourse interrupted by silence at both the longest (1.25-s) and shortest (20-ms) durations available through turning of the control dial, the subjects were allowed to briefly explore the effects of different interruption rates by turning the control dial themselves. They were then permitted to make two practice adjustments for normal rate discourse interrupted by silence and by noise. Prior to each threshold adjustment, the control knob was set to produce the highest interruption rate of 25 ips (interruptions of 20 ms). Each listener made a total of 18 formal threshold adjustments, with 6 adjustments made at each playback rate in a separate block of trials. The order in which signal rates were presented was original, slow, and fast for half of the listeners, and was original, fast, and slow for the remaining listeners. Within each block, adjustments were made alternately with silence and noise as interrupters, beginning with interpolated silence. Listeners were given as much time as needed to make their threshold adjustments. The average duration of an experimental session, including instruction and debriefing, was approximately 25 min.

II. RESULTS AND DISCUSSION

The median off-time for a listener's three judgments of the lower limit of speech discontinuity was considered the deletion detection threshold for each condition. The means of those median off-times are presented in Table I for interruption by silence and by noise for the three speech rates employed. A two-way analysis of variance for repeated measures yielded significant main effects of interrupter ($F = 61.64$, $p < 0.001$) and speech rate ($F = 27.06$, $p < 0.001$) and a significant interaction ($F = 15.68$, $p < 0.001$). Subsequent Tukey tests indicated that thresholds were significantly higher ($p < 0.01$) at each signal rate when speech was interrupted by noise rather than silence. Thresholds also differed across playback rates ($p < 0.01$) for all comparisons) when the speech stimuli were interrupted by noise. By Tukey tests, thresholds did not differ across playback rates when the speech stimuli were interrupted by silence. However, Dunnett comparisons for the silent gap conditions did indicate ($p < 0.01$) that interruption thresholds were higher at the decreased playback rate than at the remaining rates.[1]

TABLE I. Mean deletion detection thresholds (in ms off-time) for connected discourse at three playback rates. Changes in thresholds from values obtained at the normal rate are given as Δ%.

	Playback rate		
Interrupter	Normal	Increased 15%	Decreased 15%
	Mean	Mean (Δ%)	Mean (Δ%)
Noise	278.7	229.4 (− 17.7)	312.7 (+ 12.2)
Silence	99.3	92.0 (− 7.4)	114.0 (+ 14.8)

The interpolation of noise rather than silence in the speech-free portions of the switching cycle produced an increase in threshold gap durations, this increase ranging from 137 ms at the most rapid speech rate to about 199 ms at the slowest rate. The resulting upper durational limits of discourse continuity with interpolated noise (ranging from about 230 ms to about 313 ms off-time, depending on signal rate) are similar to the durational limits previously observed by Bashford and Warren (1987) using the same interruption paradigm but a different discourse passage.

In that earlier study, as mentioned briefly in the Introduction, deletion detection thresholds were obtained for a recorded excerpt read from an article appearing in a popular magazine. When regularly spaced gaps in the verbal stimulus were filled with spectrally matched noise, threshold gap durations averaged 304 ms, corresponding to 99% of the average duration of words in that passage (306 ms). Similar effects of interpolated noise were obtained with the passage employed in the present study. When the reading of the Constitution was played back at its original recording speed, threshold off-time with interpolated noise corresponded to 94% of the average word duration. When the speed of playback was increased, so as to temporally compress all components of the signal by 15%, threshold off-time decreased by 17.7%, with the duration of periodic gaps equal to 91% of the average word. In contrast, when the speed of playback was decreased to produce a 15% expansion of the speech signal, the threshold off-time increased 12.2% and equaled 92% of the average word duration. The average percentage of shift in threshold durations produced by a 15% rate change, disregarding the direction of change, was 14.95%.

Thus, as measured in the present study and in the earlier experiments of Bashford and Warren (1987), the discontinuity threshold appears to reflect an informational limit rather than a fixed temporal limit for verbal induction of discourse. Because the interruptions employed in these studies were not linked systematically to specific speech components, interpretation of the results in terms of possible sampling limits for perceptual restoration must be considered as statistical. On average, induction through regularly spaced interruptions begins to fail when gaps in the speech signal approximate word length: Under these conditions, listeners would receive only a single fragment of an average length word. However, as mentioned in the Introduction, Bashford and Warren found that threshold gap durations dropped from about 300 ms for a normal reading of discourse to about 150 ms (half the average word duration) for the same discourse passage read at the same rate but with the order of words reversed, and a similarly low threshold duration was obtained for isolated monosyllables. Thus it appears that the possible "world-length" limit for phonemic restoration does not apply when suprasegmental syntactic and semantic context is absent.

ACKNOWLEDGMENT

This study was supported by grants from NIH and AFOSR.

[1] In the earlier experiments of Bashford and Warren (1987), interruption thresholds varied dramatically for different types of speech when gaps were filled with noise but were equivalent across stimuli (threshold gap duration about 50 ms) when interpolated silence was employed. Interruption thresholds in the silent gap conditions of the present study were higher than in those previous experiments and also appear to have been influenced to some extent by temporal properties of the speech signals. This difference in results is probably attributable to a change in the instructions given listeners in the present study. Participants in the earlier experiments experienced greater difficulty making threshold judgments with interpolated silence than with interpolated noise. Speech signals interrupted by noise typically appear to be both continuous and natural up to a listener's threshold off-time, and, beyond that threshold duration, detectable gaps appear relatively large. In contrast, speech interrupted by silence appears unnatural at all interruption rates. Even with very brief silent gaps, the speech signal has a "rough" or "bubbly" quality and listeners may spend considerable time attempting to make judgments within the fairly wide range of rapid interruption rates producing that effect. In the present study, an attempt was made to simplify the judgments required of listeners: They were instructed to base their adjustments upon the production of detectable gaps, and to avoid judgments based on roughness. As anticipated, threshold gap durations with interpolated silence were greater, while thresholds for interruption with interpolated noise appear to have been unaffected by this change of instructions.

Bashford, J. A., Jr., and Warren, R. M. (1987). "Multiple phonemic restorations follow the rules for auditory induction," Percept. Psychophys. **42**, 114–121.

Houtgast, T. (1972). "Psychophysical evidence for lateral inhibition in hearing," J. Acoust. Soc. Am. **51**, 1885–1894.

Verschuure, J. (1978). "Auditory excitation patterns," Doctoral dissertation, Erasmus University Rotterdam, The Netherlands.

Warren, R. M., and Obusek, C. J. (1971). "Speech perception and phonemic restorations," Percept. Psychophys. **9**, 358–362.

Warren, R. M., Obusek, C. J., and Ackroff, J. M. (1972). "Auditory induction: Perceptual synthesis of absent sounds," Science **176**, 1149–1151.

Warren, R. M. (1984). "Perceptual restoration of obliterated sounds," Psycholog. Bull. **96**, 371–383.

Appendix E

Perception of complex tone pairs mistuned from unison

Richard M. Warren, James A. Bashford, Jr., and Bradley S. Brubaker
University of Wisconsin—Milwaukee, Department of Psychology, Milwaukee, Wisconsin 53201

(Received 10 June 1988; accepted for publication 27 March 1989)

Periodic sounds mistuned from unison may interact to produce pitch glides: When a broad-spectrum complex tone having a fundamental frequency of 400 Hz or less and containing several harmonics above the 8th is mixed with itself after a slight change in the waveform repetition frequency (1 Hz or less), listeners hear a rising glissando when corresponding portions of the waveforms approach alignment and a falling glissando as they recede from alignment. Glissandi are unimpaired if harmonics below the 8th are absent, but if, instead, harmonics above the 8th are removed, only amplitude fluctuations are heard (not glissandi). When two broad-spectrum complex tones with independent, randomly derived phase spectra are mistuned slightly from unison and mixed, complex repeated patterns other than glissandi are heard. These observations, along with others involving a variety of periodic sounds mistuned from unison, provide information concerning the nature of frequency domain and time domain mechanisms employed for the perception of iterated acoustic patterns.

PACS numbers: 43.66.Hg, 43.66.Ki [NFV]

INTRODUCTION

A number of interesting phenomena can be observed when complex tones are mistuned slightly from unison (frequency ratio of 1:1).

When a pair of mistuned complex tones having fundamental frequencies that differ by Δf_0 are mixed, their corresponding harmonics (that is, those with the same harmonic number n) beat at a rate of $n\Delta f_0$. The beat rates thus form a harmonic series consisting of integral multiples of Δf_0. It has been reported that, under some conditions, these harmonically related beats are integrated perceptually to form a complex pattern of amplitude fluctuations heard to repeat at the same rate as the beats produced by the spectral fundamentals whether or not the fundamentals are present (Warren, 1978, 1982, pp. 100–101).

For some purposes, it can be useful to consider the effects of mistuning complex tones from unison in terms of waveform interactions rather than the interactions of spectral components. Consider two broad-spectrum complex tones identical in every way (the same waveform and the same amplitude and phase spectra). If the waveform of one of these complex tones is stretched slightly (so that its repetition frequency drops by a small amount), and these two "correlated" tones with matching phase spectra are then mixed, the temporal separation of corresponding portions of the waveforms of the component tones will be continually changing, with alignment occurring at a rate equal to the difference in frequency of the complex tones (Δf_0). It is known that when a complex sound is mixed with itself following a delay of τ seconds, a "repetition pitch" of $1/\tau$ Hz may be heard (for review, see Plomp, 1976, pp. 138–139). Since the mistuned correlated tones have continually changing displacements from synchrony, pitch glides might be heard: As their waveforms move away from alignment, a downward gliding pitch might be produced by the increasing value of τ, followed by a rising pitch glide as the waveforms move past the point of maximum separation and back toward alignment. Such glissandi have been reported for pairs of pulsate periodic stimuli mistuned from unison (Thurlow and Small, 1955), and the changing pitch was attributed to the changing temporal separation of the discrete pulses (see also Small and McClellan, 1963). However, the present study demonstrates that these glides are not restricted to mistuned periodic pulses, but are found for the more general case of mistuned nonpulsate complex tones.[1] It should be emphasized that these glissandi would be anticipated only for the mistuning of complex tones with matched or nearly matched phase spectra. When members of a pair of mistuned complex tones have waveforms with independent, randomly derived phase spectra for corresponding harmonics, there can be neither alignment nor delay from alignment of corresponding portions of the component waveforms, and hence no pitch glide would be expected.

I. PRELIMINARY OBSERVATIONS

A number of informal observations were made by a panel of four psychoacoustically experienced subjects who listened to the interaction of two broadband complex tones that had been mistuned slightly from unison and then mixed. These complex tones were presented diotically through headphones at levels ranging from 30–60 dB SPL, and consisted of all harmonics up to 8 kHz. All tone pairs had one member with a randomly determined phase spectrum, while the other member of the pair was either "correlated" (identical except for a slight difference in frequency) or "uncorrelated" (having a slight difference in frequency and independent randomly determined phases for corresponding harmonics). Details of the manner in which such complex tone pairs were produced will be given subsequently in Sec. II. Since these preliminary observations formed the basis for the formal study, they will be listed below.

A. Correlated complex tone pairs mistuned from unison

(1) Glissandi became less clear with increasing fundamental frequencies of the tone pairs, and were judged to be absent above roughly 400 Hz.

(2) There was no lower frequency limit for tone pairs producing glissandi, and pitch glides were also heard clearly for mistuned complex waveforms having infratonal repetition frequencies (i.e., repetition frequencies below 20 Hz).

(3) Glissandi were not heard if the extent of mistuning exceeded a critical value ("chirps" were reported instead). This limit for perception of a gliding pitch was a discontinuous function of the fundamental frequencies of the complex tones.

(4) Glissandi required the presence of higher but not the lower harmonics of the complex tones. When tone pairs with fundamental frequencies from 50–400 Hz were high-pass filtered above the 7th harmonic, perception of pitch glides appeared unimpaired. However, glissandi were not heard when the same tone pairs were low-pass filtered below the 8th harmonic (complex patterns of amplitude fluctuation were heard instead).

(5) Removal of even-numbered harmonics from mistuned broadband tone pairs doubled the glissando rate, so that two pairs of rising and falling pitch glides were heard from one waveform alignment to the next.

B. Uncorrelated complex tones mistuned from unison

(1) Pitch glides could not be perceived, but complex patterns of amplitude fluctuation were heard to repeat at Δf_0 under some conditions. Thus, when 198- and 200-Hz uncorrelated tone pairs were mixed, a clear pattern repeating at 2 Hz was reported. Removal of f_0 (the spectral fundamental) of each tone did not change either the clarity or ensemble repetition rate of these complex patterns.

(2) The beating of individual harmonic components of the complex tones was difficult to resolve when Δf_0 was 1 Hz or more. But, when the uncorrelated tones were mistuned by less than about 0.5 Hz, the integrated or ensemble periodicity became less salient and individual beat rates were clearly dominant.

(3) Removal of the even-numbered harmonics sometimes resulted in hearing ensemble periodicities of $2\Delta f_0$ as well as Δf_0.

II. GENERAL METHODS

A. Subjects

Listeners were all trained in psychoacoustic experimentation. Informal observations were based upon reports of between four and six listeners who served as observers for each of the phenomena described, and all of the phenomena were heard by each of these listeners unless otherwise stated. The formal experiments employed listeners drawn from this pool.

B. Stimuli and apparatus

Two types of complex tones were employed. "Frozen noise tones" were generated by the repetition of waveforms excised from a 100- to 8000-Hz band of pink noise. These tones had randomly determined amplitudes and phases for harmonic components which extended from the fundamental frequencies up to 8000 Hz (for a discussion of repeated noise segments as model periodic stimuli, see Warren, 1982, pp. 78–80). "Synthesized" complex tones consisted of all harmonics lying between the fundamental frequency and 8 kHz, with each harmonic having the same amplitude and an individually specified, randomly determined phase. These tones were generated from polynomial equations by a Data Precision Co. polynomial waveform synthesizer model 2020-100 (512 000/16-bit data point capacity, operated at a sampling frequency of 50 kHz) in conjunction with a Hewlett–Packard model 9816 computer.

Two matched factory-modified digital delay lines (Eventide model BD955) were used in the production of all frozen noise tones. When placed in recirculating or "looped" mode, the delay lines repeated stored input without change. The delay lines (operated in conjunction with appropriate antialiasing and reconstruction filters) had a flat frequency response (± 1 dB) from 50–16 000 Hz and a 60-dB dynamic range, based upon a sampling frequency of 50 kHz and 10-bit coding. The storage times (and repetition periods) used in the present study ranged from 2.5 ms–1 s. To produce the correlated periodic stimuli mistuned from unison, the same waveform (either a noise segment or a synthesized waveform) was repeated at identical repetition frequencies, and then one of these identical waveforms was temporally stretched by lowering the 8-MHz clock frequency by the desired amount. When the experiment required the halting or freezing of a glide at some particular delay, the changed clock frequency was returned to the original value when the desired asynchrony of corresponding portions of the two waveforms was reached. For the uncorrelated tones based on frozen noise, members of a tonal pair were derived from separate segments of a noise, and for the synthesized uncorrelated tones derived from polynomial equations, a different set of randomly determined phases was assigned to the equations for the corresponding harmonics of a complex tone pair. Two Hewlett–Packard model 3325A function generators locked to the same time base controlled the sampling frequencies (nominally 50 kHz) for all stimulus pairs. These generators were adjustable in steps of 1 mHz, and changes in the repetition period of the stored waveforms were executed by a Hewlett–Packard model 85 computer which controlled the frequency of one of the delay line clocks. The outputs of the delay lines were combined using an audio mixer.

Spectra of stimuli were monitored with a Brüel and Kjaer model 2033 spectrum analyzer, and waveforms were monitored with a two-channel digital storage oscilloscope (Nicolet model 3091). Sound spectrograms were generated by a Kay model 7800 digital sonagraph.

When desired, high- and low-pass filtering of the mixed iterated waveforms was accomplished with Wavetek/Rockland filters, either model 751A having attenuation slopes of 115 dB/oct, or model 852 with slopes of 48 dB/oct.

Stimuli involving repeated frozen noises were generated on-line during the experiments, while stimuli involving synthesized waveforms were prerecorded on an Otari model MTR 90-II 16-track recorder.

C. General procedure

Subjects were seated in an audiometric room, and the stimuli were delivered diotically through a matched set of headphones (either TDH-49 or Sennheiser HD 230) at a comfortable level selected by the listener (usually between 30- and 60-dB SPL), unless otherwise noted. Use of higher stimulus levels favored the hearing of beats, which could distract listeners from attending to the gliding pitch.

III. EXPERIMENT 1: LIMITS OF MISTUNING PRODUCING PITCH GLIDES

Preliminary experiments using complex correlated tones had indicated that the maximum deviation from unison for which pitch glides could be heard was a discontinuous function of the fundamental frequencies of the tone pairs. The present experiment confirmed those observations: The limit of mistuning was found to be proportional to fundamental frequency for tone pairs of from 1 to 25 Hz, and was found to be a constant value of about 1 Hz for tone pairs having fundamental frequencies of from 50 to 400 Hz. As will be discussed below, the results obtained for both ranges appear to follow the same simple rule.

A. Stimuli and procedure

The fundamental frequencies of the correlated pairs of frozen noise tones before mistuning from unison were 25, 50, 100, 200, and 400 Hz. In addition, correlated frozen noise waveforms repeated at infratonal frequencies of 1, 2, 4, 8, and 16 Hz were employed.[2] Each pair of mistuned periodic stimuli was derived from a single randomly selected noise segment of desired duration which was repeated on both delay lines. The mistuning from unison was in steps which varied with the initial iteration frequency. For repetition periods from 1 s–20 ms, these steps were integral multiples of 0.5% of the initial values, and for periods from 10–2.5 ms, steps were multiples of 0.1% (as described earlier, mistuning was accomplished through a computer program controlling the clock frequency driving one of the delay lines).

Subjects were instructed to report hearing glissandi only when a gliding pitch was clearly perceived. When mistuning exceeded the limit for hearing rising and falling pitch glides, listeners reported hearing repetitions of a single brief sound resembling a chirp. They received five blocks of 20 trials; each block involved all ten repetition frequencies (ranging from 1–400 Hz), which were presented in random order. Every repetition frequency within a block was judged twice using the method of limits. For the first of these judgments, the experimenter mistuned one frequency beyond the point where glissandi could be heard (the initial mistuning was randomly selected from the range of 5–12 steps of the size described above). Mistuning was decreased stepwise until glissandi were heard. The experimenter then decreased the extent of mistuning further by an additional two to four units (subjects always indicated that they heard glissandi for this extent of mistuning). The extent of mistuning from unison was then increased systematically, using steps of the size described above, until the subject reported that a glissando could not be heard. The upper limit for hearing glissandi on that pair of trials was considered to be the average of the ascending and descending orders of presentation. A particular frozen noise segment was used for only one pair of trials for a single subject.

For noise segment repetition frequencies of 1 and 2 Hz, the minimum time required for a pitch glide period (a pair of rising and falling glissandi) was inconveniently long—therefore, only the falling pitch glide following initial waveform alignment was judged, and each frozen noise segment was used for only one judgment.

B. Results and discussion

As shown in Fig. 1, glissandi could not be heard for mistuned correlated tone pairs from 50–400 Hz if the difference in fundamental frequencies (Δf_0) was greater than about 1 Hz. In contrast, for stimulus pairs with waveform repetition frequencies from 25 down to 1 Hz, the limiting value of Δf_0 for pitch glides was proportional to the repetition frequencies of the pairs of periodic sounds, and a linear function ($R = 0.996, p < 0.0001$) was obtained.

The acoustic interactions of the harmonics of complex tones mistuned from unison are illustrated in Fig. 2. The sound spectrograms show the amplitude changes produced by the beating harmonics for 100- and 99.5-Hz correlated tones (top spectrogram) and for 100- and 99.5-Hz uncorrelated tones (bottom spectrogram). It can be seen that the rate of beating of corresponding spectral components (i.e., components of the mixture having the same harmonic numbers) is equal to the fundamental beat rate (0.5 Hz) multiplied by the harmonic number whether the tone pair is correlated or uncorrelated. Figure 2 also shows that, for a correlated tone pair, beating harmonics all have their amplitude maxima occurring simultaneously once each 2 s at the alignment of the corresponding portions of the two waveforms. The upward sweeping pattern of the spectrogram is produced as the waveforms move toward alignment and the

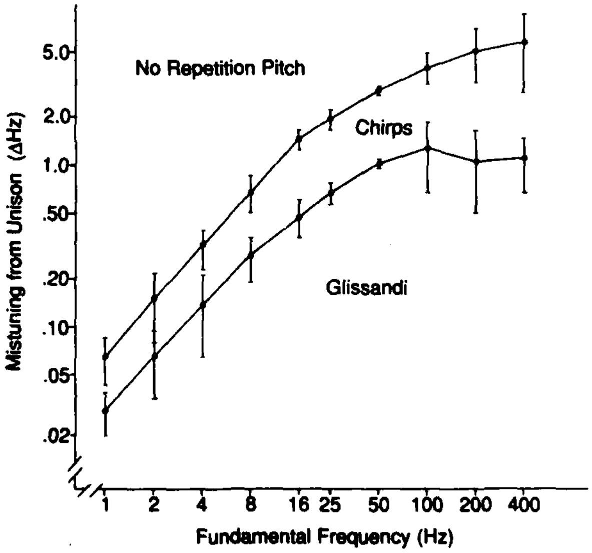

FIG. 1. Means and 95% confidence intervals for glissando and chirp limits: The greatest mistuning from unison permitting perception of glissandi (experiment 1) and chirps (experiment 2). The frequency of one member of the pair is given by the abscissa, and the decrease in frequency of the second periodic sound (produced by stretching the waveform of the first) is given by the ordinate. For further details, see text.

downward sweeping pattern as the waveforms move away from alignment. The sound spectrograms exhibit the same resolution in hertz for the upper and lower harmonics (the sound spectrograms shown in Fig. 2 are based upon a fixed filter bandwidth of 150 Hz). However, the resolving power of our auditory system is rather different from that of a spectrograph, with auditory resolution in hertz decreasing with increasing harmonic number (see Plomp, 1976, pp. 1–25). As will be shown in experiment 3, glissandi require the presence of unresolved upper harmonics but not resolvable lower harmonics.

Despite the discontinuity of the function for glissandi shown in Fig. 1, a simple rule applies over the entire range: *Glissandi cannot be heard unless the individual glide durations last at least 0.5 s.* Let us consider first how this rule applies to the range of fundamental frequencies from 50–400 Hz and results in the horizontal boundary separating glissandi from chirps. This horizontal segment corresponds to a pitch–glide cycle repeated at a frequency of about 1 Hz (period of 1 s), with the temporally contiguous upward and downward glides each lasting 0.5 s. Adding a replica of a sound to itself after a delay of τ seconds produces a repetition pitch corresponding to a tone of $1/\tau$ Hz (see Plomp, 1976, p. 139). Glissandi may be considered as gliding repetition pitches. The maximum repetition delay corresponds to one-half the tonal period, resulting in a lower pitch limit of the glide 1 oct above the pitch of the fundamental tones. Despite the changes in this lower pitch limit from 800 Hz (for the 400-Hz tonal pair) to 100 Hz (for the 50-Hz tonal pair), the limiting duration of each glide remained at 0.5 s, so that it was glide time rather than rate or extent of pitch change that determined the limit of mistuning for glissandi. Figure 1 shows that, for fundamental frequencies from 1–25 Hz, the mistuning limit for hearing glissandi was a straight line function having a positive slope. Calculations show that throughout this range, the glide time required to reach a waveform asynchrony of 10 ms (repetition pitch of 100 Hz) remained constant at 0.5 s. At fundamental frequencies of 50 Hz and above, the end of a downward glide and the start of the upward glide were temporally contiguous, whereas a temporal gap separated these glides for fundamentals below 50 Hz.

IV. EXPERIMENT 2: LIMITS OF MISTUNING PRODUCING CHIRPS

When mistuning from unison in experiment 1 exceeded the limit for producing glissandi, the separate upward and downward pitch glides merged into a single percept—that of a transitory chirp. Further mistuning resulted in a loss of the chirplike quality, and an iterated amplitude pattern was heard having a period equal to the difference between the fundamental frequencies of the periodic sounds.

The present experiment measured the maximum extent of mistuning for which chirps could be heard. The experimental subjects were those who had previously served in experiment 2.

A. Stimuli and procedure

The stimuli and the experimental procedure were the same as in experiment 1—the only difference was in the task. Listeners were instructed to judge the upper limit of mistuning for which chirps could be heard (greater mistuning re-

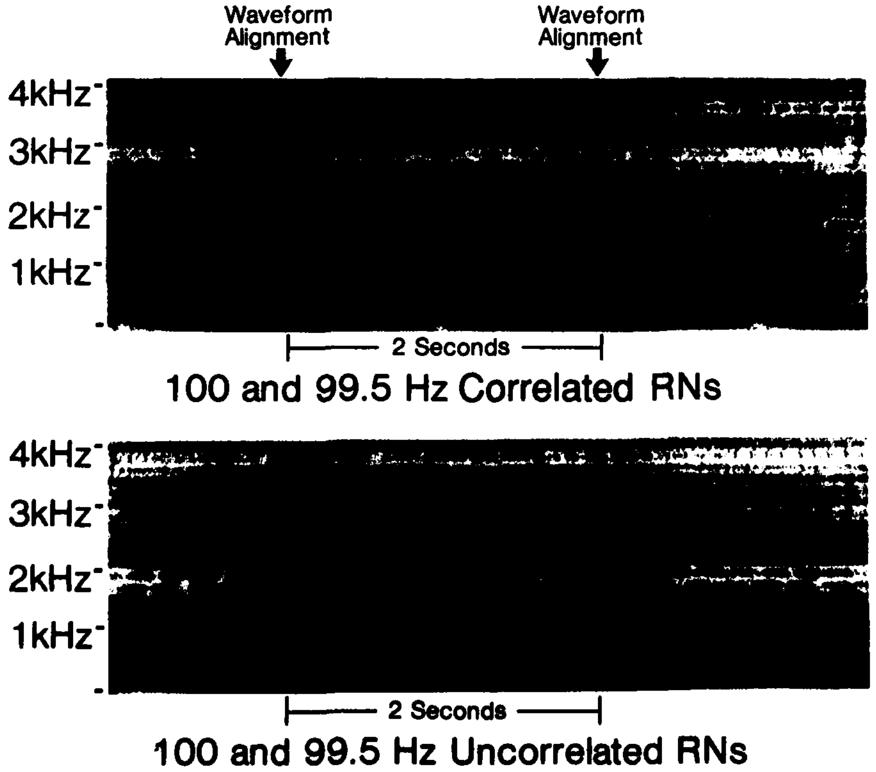

FIG. 2. Sound spectrograms of mixtures of 99.5- and 100-Hz reiterated noises (RNs). The top spectrogram is based upon correlated waveforms (both tones are derived from the same 10-ms segment of Gaussian noise, with the 99.5-Hz tone produced by a 0.5% "stretching" of the waveform). The bottom spectrogram is based upon uncorrelated waveforms (independent segments of Gaussian noise). Waveform alignment occurs only for the top spectrogram.

sulted in perception of a periodic amplitude fluctuation lacking any pitchlike or chirplike quality).

B. Results and discussion

As shown in Fig. 1, the mistuning limit for perceiving chirps is between two and three times the mistuning limit for glissandi occurring with the 1- through 100-Hz fundamental frequencies, increasing to about five times the glissando limit for the 200- and 400-Hz fundamental frequencies.

Both glissandi and chirps appear to be based upon repetition pitch, and the upper mistuning limits for chirps seem to reflect the rate-of-change limit for this type of pitch. Tones having rapidly changing frequencies also give rise to chirps, but the bases and characteristics of tonal chirps and repetition pitch chirps are quite different (for descriptions and discussions of tonal chirps, see Náběelek *et al.*, 1970, 1973; Trenque and König, 1988).

V. EXPERIMENT 3: HARMONIC COMPONENTS REQUIRED FOR PITCH GLIDES

Informal observations indicated that resolved harmonics heard alone do not produce glissandi. For example, when 99.5- and 100-Hz broadband tones derived from the same frozen noise segment were mixed and then low-pass filtered at the 6th harmonic (filter slopes of 115 dB/oct), glissandi could not be heard, and only periodic amplitude fluctuations produced by the beating lower harmonics were evident. In order to confirm this observation and to determine more accurately the low-pass threshold for glissandi, correlated complex tones were synthesized from polynomial equations.

A. Stimuli and procedure

The polynomial waveform synthesizer was used to generate a series of complex tones with fundamental frequencies of 100 Hz. The members of this series consisted of all harmonics up to the 4th, 6th, 8th, 10th, 12th, and 14th, respectively. Each of these complex tones had harmonics of equal amplitude and different randomly assigned phases, and was used to produce a correlated tone pair. A second series of complex tone pairs was synthesized which differed from the first series only in having different random assignments of phase. The complex tones of each series were used to produce tone pairs mistuned from unison with fundamentals of 100 and 99.8 Hz, using the procedure described in Sec. II. The six listeners judged whether glissandi were heard. The method of limits was used, starting with the correlated tones having 14 harmonic components (which always produced glissandi) and continuing with complex tones having successively decreasing numbers of harmonics until glissandi were no longer heard. Judgments were then made starting with complex tones having only four harmonics (for which glissandi were never heard) and continuing with complex tones having successively increasing numbers of harmonics until glissandi were reported. After the first pair of judgments, the highest harmonic for the decreasing series was randomly selected as 14 or 12, and for the increasing series as 4 or 6. After three trials (pairs of judgments), an additional three trials were run with the second series of complex tones having different randomly assigned phases. The threshold value for each trial was the average for the descending and ascending modes of presentation.

B. Results and discussion

Table I gives the thresholds obtained with six subjects for each of the complex tones. It can be seen that, on the average, harmonics above the eighth must be present for glissandi to be heard.

Commenting on a preliminary report by Warren *et al.* (1984) dealing with glissandi produced by tones mistuned from unison, Hartmann (1985) suggested that the pitches we described were caused by the effects of a "frequency domain grating" analogous to the effects of diffraction gratings used in optics. The glides were attributed to the orderly progression of spectral maxima, as if a filter were swept through the harmonics of the complex tones. It was considered that not all harmonics entered into pitch glides: In the case of tones consisting of the first eight harmonics, only four would enter into the orderly progression of maxima, and with such a small number of harmonics forming a progression, glides would be difficult to hear.

If the theory described above were valid, the pitches heard would be limited to the range covered by harmonic components. In order to determine if glissandi can reach pitches outside the range of spectral components, we synthesized mistuned correlated tones, each consisting of the 71 harmonics from the 10th through the 80th. One member of the pair had a (missing) fundamental of 100 Hz with harmonics of equal amplitude and randomly assigned phases. The other member of the pair was derived from the first, and had a (missing) fundamental of 99.95 Hz. When mixed, a glissando pair was produced having a period of 20 s with a falling pitch lasting 10 s and a rising pitch of the same duration. Of especial interest was the fact that, at the transition from falling to rising pitch, all odd-numbered harmonics were canceled and a new complex tone was formed consisting of the 5th through the 40th harmonics of 200 Hz. A pitch corresponding to the new (missing) fundamental of 200 Hz was heard clearly at this point, and this pitch appeared to be continuous with both the falling and rising segments of the glissandi for all six listeners. Since the spectral domain grat-

TABLE I. Number of successive harmonics for each complex tone (starting with the fundamental) required for perception of glissandi with correlated tones of 100 and 99.8 Hz (for further description, see the text).

	Trial number						
Subject	1	2	3	4	5	6	Average
JB	8	8	8	9	9	10	8.66
BB	11	11	11	9	9	10	10.16
PK	10	11	9	11	9	9	9.83
JR	8	8	7	6	6	6	6.83
DT	7	6	8	6	7	7	6.83
RW	8	7	9	8	7	7	7.67
Combined	8.67	8.50	8.67	8.17	7.83	8.17	8.33

ing theory would limit the range of pitches heard to those of the harmonics actually present, it would seem that this particular spectral explanation for glissandi does not apply.[3]

We have seen that mistuned correlated complex tones do not produce glissandi if harmonics above the 8th are absent. Spectral resolution of harmonic components can be accomplished by listeners up to about the 8th (Plomp, 1964; Plomp and Mimpen, 1968), so the presence of unresolved harmonics may be necessary for glissandi. Alternatively, glissandi may require more than eight harmonics for each member of the correlated tone pair. In order to test these hypotheses, a complex tone was synthesized consisting of eight harmonics of 100 Hz from the 10th through the 17th, each harmonic having the same amplitude and a different randomly determined phase. This complex tone was used to generate a second correlated tone with a fundamental frequency of 99.8 Hz. All six listeners heard a pair of rising and falling pitch glides repeated each 5 s, suggesting that the lack of glissandi for correlated tones consisting of harmonics from the fundamental through the 8th is attributable to their spectral resolution rather than to the small number of components.

In physical acoustics, spectral domain and temporal domain analyses may be equally valid and interchangeable—but, in physiological acoustics, each of these analyses is associated with different modes of neural processing. What then are the auditory bases for hearing glissandi? Both temporal and spectral mechanisms can be formulated. First, a spectral basis: As the correlated waveforms move out of alignment, the resulting peaks and troughs in the spectral envelope provide corresponding place cues on the basilar membrane. With the initially small asynchronies, the interpeak spacing is large and the pitch heard (which corresponds to the frequency separation of adjacent peaks) is high. As asynchrony becomes larger, the interpeak separation becomes smaller, and the pitch drops until all odd harmonics are canceled at the point of maximum asynchrony. The pitch then rises until alignment is reached once more, and the cycle then repeats. A possible temporal basis involves the complex periodic patterns of amplitude fluctuations produced at individual loci on the basilar membrane by the interactions of several pairs of unresolved harmonics. An autocorrelational analysis may lead to a pitch equivalent to the reciprocal of the repetition period of these complex patterns (for a detailed discussion of repetition pitch based upon the autocorrelation of neural patterns, see Yost *et al.*, 1978).

VI. EXPERIMENT 4: MATCHING OF PITCHES ASSOCIATED WITH GLISSANDI

In order to determine if the glissandi produced by tones mistuned from unison were the consequence of repetition pitch, glides were halted at particular waveform asynchronies by reestablishing unison, and listeners then matched the terminal pitch of the glide with that of a sinusoidal tone. If glissandi represent changes in repetition pitch, then it should be possible to match the pitch of a glide halted at an asynchrony of τ s with that of a sinusoidal tone having a frequency of $1/\tau$ Hz.

Pitches corresponding to static waveform asynchronies are less salient than gliding pitches; however, halting a glide at a particular pitch can highlight the steady-state value. Informal observations indicated that pitch matching was facilitated more by a rising than a falling pitch, and so all asynchronies employed in the present experiment were reached by freezing an upward gliding pitch. Experiment 3 has demonstrated that lower harmonics are not necessary for hearing glissandi, and, in order to eliminate the possibility that the pitches of these resolvable harmonics would interfere with static repetition pitch judgments, they were removed by filtering.

A. Stimuli and procedure

All complex tones were synthesized from polynomial equations. Following bandpass filtering, the tones consisted of all harmonics from the 11th up to 8000 Hz. Before mistuning from unison, each member of the tone pair had the same (missing) fundamental frequency, which was either 50, 100, or 200 Hz. As described in Sec. II, the phase spectrum of each of the synthesized tone pairs was randomly determined, and all of the harmonics had the same amplitude. The six terminal waveform asynchronies of the glides used for pitch matching corresponded to theoretical values of repetition pitches ($1/\tau$ Hz) of 500, 600, 700, 800, 900, and 1000 Hz, respectively. A block consisted of six trials with the same fundamental frequency (50, 100, or 200 Hz) involving each of the six values of τ presented once in a randomly determined order. Each trial began with a maximum possible asynchrony of the correlated waveforms ($1/2f_0$ s), and the extent of mistuning was varied randomly so that the time which would be required to reduce the asynchrony to 1 ms (corresponding to a repetition pitch of 1000 Hz) would lie within the range of 24–31 s. However, the asynchrony was not always allowed to reach this value. When τ corresponded to one of the terminal asynchronies described above, the glide was halted and the listener matched the terminal pitch by adjusting the frequency of a sinusoidal tone. Adjustment was accomplished by turning a dial on a frequency synthesizer having dial markings concealed from view. Listeners could switch from the complex tone pair to the sinusoidal tone at will.

Before starting the experiment, listeners practiced pitch matching with tone pairs of 45 Hz (which differed from the fundamental frequencies used in the formal experiment), and terminal delays were employed corresponding to repetition pitches of 450, 540, 630, 720, 810, and 900 Hz. The listener received information concerning both the stimulus repetition pitch and their choice of matching pitch after each trial was completed. During training with the 45-Hz fundamental tones, as well as during the formal experimental session with the 50-, 100-, and 200-Hz fundamental tones, the listener could elect to: (1) hear the pitch glide to the terminal value again (up to four times); (2) go on to the next trial without deciding on a pitch match and returning to that frequency later; or (3) hear a single sweep that covered the entire range of experimental repetition pitches using the same fundamental frequency but with a different waveform and glide rate—this last option aided in avoiding the octave errors that frequently occur in pitch matching experiments. After completing three to five practice blocks, listeners indicated that they were sufficiently familiar with the task, and the formal experiment began.

Subjects completed five blocks of six judgments at each of the three fundamental frequencies (50, 100, and 200 Hz) for a total of 90 judgments per subject. For a given fundamental frequency, each of the five blocks of judgments employed a tone pair with a different, randomly determined phase spectrum. Subjects completed all judgments at the 50-Hz fundamental frequency before proceeding to fundamentals of 100 Hz and, finally, 200 Hz. No knowledge of results was provided until all judgments at a given fundamental frequency were completed. Because the static repetition pitches were somewhat weaker for the 100- and 200-Hz fundamentals, formal data gathering with those stimuli did not begin until a subject's six judgments within a block reached a criterion of $\pm 5\%$ of $1/\tau$ Hz (listeners were not informed when this criterion was reached). Little practice was required: Two of the three subjects reached criterion during the first block of trials with both the 100- and 200-Hz fundamentals. The third subject required two blocks of trials with the 100-Hz fundamental and six blocks with the 200-Hz fundamental.

B. Results and discussion

While the observations in experiments 1 and 2 were consistent with the hypothesis that glissandi represent changes in repetition pitch, it was not established that the gliding pitches perceived for the mistuned tone pairs had the requisite values.

Figure 3 shows that a glissando frozen at a waveform asynchrony of τ seconds had a pitch of $1/\tau$ Hz as required by the repetition pitch hypothesis. This agreement held for all six values of τ employed with each of the three tonal fundamental frequencies (f_0's) (50, 100, and 200 Hz). Standard errors of the mean (SEs) were calculated for matches at each value of $1/\tau$ at each f_0 collapsed across listeners: Discounting the single judgment that involved an octave error, SEs were less than 6 Hz for $1/\tau$ values of 500 and 1000 Hz, and were no more than 2 Hz at the four other values of $1/\tau$.

VII. EXPERIMENT 5: GLISSANDO DOUBLING FOR ODD-HARMONIC COMPLEX TONES

Listeners heard a clear doubling of the rate of glissandi when even-numbered harmonics of correlated complex tones were absent. Such odd-numbered harmonic complex tones produced pairs of rising and falling pitch glides at rates of $2\Delta f_0$ Hz so that, for example, mistuning odd-harmonic correlated tones of 199.5 and 200 Hz produced a pair of rising and falling glides repeated once per second.

Waveforms of tones consisting of only odd harmonics have a special feature: The second half of the periodic waveform is always an antiphasic (polarity inverted) repetition of the first. As the waveforms of mistuned broadband odd-harmonic complex tones move from one cophasic alignment of corresponding portions of the waveforms to the next, they also move through an antiphasic alignment halfway through this cycle. Pitch glides are associated with both types of alignment. Starting with cophasic simultaneity, a falling glissando is heard for the first quarter of the period, and a rising glide for the second quarter. At the end of the second quarter, an antiphasic superpositioning occurs, and acoustic cancellation takes place. Following this brief silent interval, a falling glide is heard for the third quarter, and a rising glide for the fourth quarter, which ends in cophasic realignment.

These observations are consistent with the hypothesis that, if τ_0 is the asynchrony from cophasic alignment of the waveforms, and τ_π the asynchrony from antiphasic alignment, then a cophasic repetition pitch is heard when τ_0 is less than τ_π, and antiphasic repetition pitch is heard when τ_π is less than τ_0. Based on this hypothesis, starting with cophasic alignment at $\tau_0 = 0$, the cophasic repetition pitch should drop until $\tau_0 = \tau_\pi$ (the end of the first quarter period), the antiphasic repetition pitch rise until $\tau_\pi = 0$ (the end of the second quarter period), the antiphasic repetition pitch drop until $\tau_\pi = \tau_0$ (the end of the third quarter period), and the

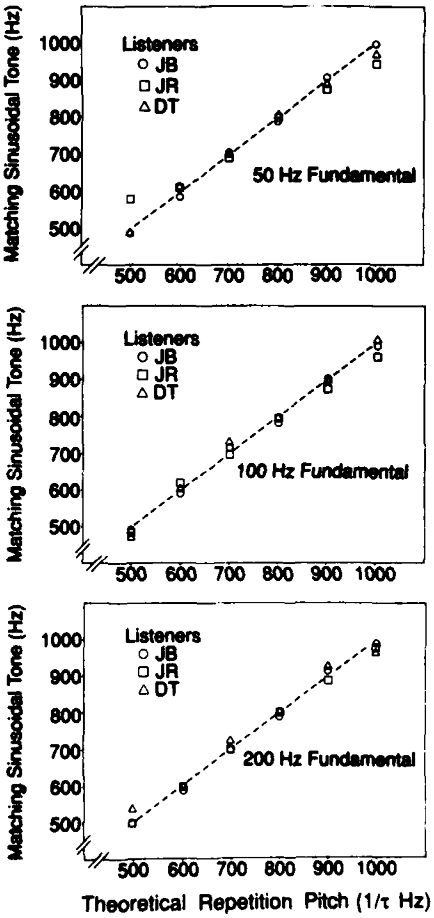

FIG. 3. Median pitch matches to glides frozen at waveform asynchronies of τ seconds. Values shown are for tone pairs having fundamental frequencies of 50, 100, and 200 Hz. The theoretical matches for repetition pitch ($1/\tau$ Hz) are given by the dashed lines. Standard errors were calculated for matches at each fundamental frequency collapsed across listeners: With the single octave error discounted, SEs were less than 6 Hz for values of 500 and 1000 Hz, and were no more than 2 Hz at the four other repetition pitches.

cophasic repetition pitch rise until the end of the cycle is reached with the return to $\tau_0 = 0$. It is well established that antiphasic repetition pitch differs from the cophasic value of $1/\tau$ Hz—antiphasic repetition pitch is ambiguous, with one value above and one below $1/\tau$ Hz. The most thoroughly investigated type of antiphasic repetition pitch is based on noise, and deviations of the antiphasic from the cophasic value of $1/\tau$ Hz are generally about $\pm 10\%$, with the exact value depending upon the value of τ and the frequency range spanned by the noise (Bilsen, 1966; Wilson, 1966; Yost *et al.*, 1978; Warren and Bashford, 1988).

The present experiment tests whether pitch glides demonstrate the predicted cophasic and antiphasic differences. Pitches were measured at appropriate waveform asynchronies for mistuned odd-harmonic complex tones mistuned from unison. The experimental procedure was similar to that employed in experiment 4—i.e., a rising glide was halted at a predetermined position, and the frozen pitch matched with a sinusoidal tone.

A. Stimuli and procedure

The synthesized complex tones mistuned from unison consisted of odd-numbered harmonics of 50 Hz, with each harmonic having the same amplitude and a randomly determined phase. As in experiment 4, the complex tones were bandpass filtered (48 dB/oct) to remove the ten lowest harmonic components and all components above 8000 Hz. One pair of mistuned correlated tones was heard with a glide starting at $\tau_0 = \tau_\pi$ ($\tau_{0,\pi} = 5 \times 10^{-3}$ s; $1/\tau_{0,\pi} = 200$ Hz), with mistuning adjusted so that the rising pitch glide (reflecting an approach toward either cophasic or antiphasic alignment) ended 11.1 s later at either $1/\tau_0 = 600$ Hz or $1/\tau_\pi = 600$ Hz. A second pair of tones was used to generate frequency glides that also started at $\tau_0 = \tau_\pi$, but ending 13.4 s later at either $1/\tau_0 = 900$ Hz or $1/\tau_\pi = 900$ Hz). Five judgments were obtained from each subject for both values of cophasic asynchrony before obtaining five judgments for both values of antiphasic asynchrony. Since judgments for τ_π should be bivalent according to the repetition pitch hypothesis, with values corresponding to roughly $0.9/\tau_\pi$ and $1.1/\tau_\pi$ Hz, trials were continued until either five judgments above or five below $1/\tau_\pi$ were obtained for each subject.

B. Results and discussion

Table II summarizes the results obtained. It can be seen that judgments were close to values of $1/\tau_0$ Hz for all listeners, while deviations from $1/\tau_\pi$ Hz varied from about 9% to 13%. While two subjects selected matches lower than the 600 and 900 Hz corresponding to $1/\tau_\pi$ Hz, one of the listeners favored higher matches. The low values for standard error scores demonstrate a consistency of matching for the individual subjects.

These results support the hypothesis that doubling of pitch glide rate when the even-numbered harmonics are absent results from the presence of antiphasic repetition pitch glides, which are not in evidence for all-harmonic tones mistuned from unison.

The spectrograms of Fig. 4 show the patterns produced by beating harmonics. Frequency sweeps resembling glissandi can be seen in these patterns, and it appears that a doubling of the sweep rate observed for mistuned all-harmonic tones is produced not only by the segregation of the odd-numbered harmonics, but also by segregation of the even-numbered harmonics. The doubling of the sweep rate for even-numbered harmonics follows from the fact that the removal of odd harmonics from all-harmonic tones produces new all-harmonic tones having double the fundamental frequency.

As discussed earlier, the resolution of harmonics by a spectrograph and by the basilar membrane is quite different. The beating of individual pairs of higher harmonics seen clearly in the spectrograms cannot be resolved by listeners, and become merged perceptually with the beating of the neighboring harmonic pairs lying in the same critical band. Although it may not be possible for a listener to resolve individual beat rates, changes in the rippling of the spectral envelope may be detected. If we imagine a narrow vertical slit slowly moving from left to right along the spectrograms of Fig. 4 (or actually view such a display with the help of a mask containing a slit), it is possible to visualize the spectral amplitude peaks sweeping down and then up in frequency as the asynchrony from waveform alignment changes with time. The frequency separation of neighboring spectral peaks decreases as the ripples sweep downward and the spacing increases with upward sweeps, in keeping with the frequency domain theories for repetition pitch (Plomp, 1976).

Why does the rate of occurrence of pitch glides drop to half when odd- and even-numbered harmonics are combined? An explanation of this apparent paradox involves the interaction of cophasic and antiphasic repetition pitches. As we have seen, an all-harmonic pitch glide falls in the first and second quarters of the period extending from one waveform alignment to the next, and a glide rises in the third and fourth quarters. When heard alone, the odd-harmonic components and the even-harmonic components each produce additional glissandi which rise in the second quarter and fall in the third. The glides based on even-numbered harmonics are cophasic, since removal of odd-harmonic components results in an all-harmonic tone of double the fundamental frequency. However, it has been demonstrated in the present experiment that the odd-harmonic glides in the second and third quarters produce antiphasic repetition pitch. When both odd and even harmonics are present, the second and

TABLE II. Mean percent deviation from $1/\tau_{0,\pi}$ Hz for pure-tone matches to the terminal pitches of cophasic (τ_0) and antiphasic (τ_π) glissandi heard with odd-harmonic complex tones. Results shown are means and standard errors (s.d./$\sqrt{5}$) for the average matches of individual listeners at two values each of $1/\tau_0$ and $1/\tau_\pi$.

	Terminal Value of $1/\tau_{0,\pi}$			
	600 Hz		900 Hz	
Subject	Cophasic ($1/\tau_0$)	Antiphasic ($1/\tau_\pi$)	Cophasic ($1/\tau_0$)	Antiphasic ($1/\tau_\pi$)
JR	+ 0.70 (0.84)	− 12.85 (1.04)	− 2.54 (0.58)	− 10.19 (0.24)
JB	− 1.55 (0.09)	− 11.33 (0.21)	− 0.28 (0.14)	− 8.76 (0.19)
DT	+ 3.33 (0.19)	+ 12.79 (0.69)	+ 3.78 (0.33)	+ 13.13 (0.22)

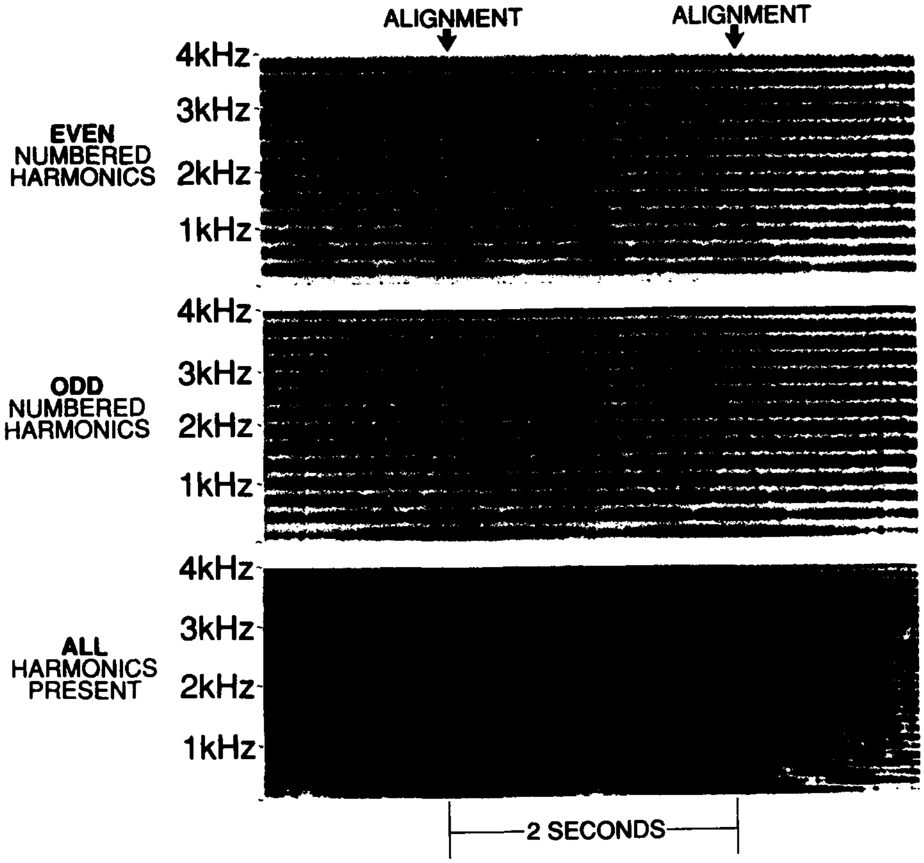

FIG. 4. Sound spectrograms illustrating the contributions of odd- and even-numbered harmonics to the acoustic interactions of complex tones mistuned from unison. The bottom spectrogram shows the beating of the components of 249.5- and 250-Hz correlated complex tones having all harmonics (odd and even) at equal amplitude and randomly assigned relative phases. The patterns produced by the beating pairs of odd- and even-numbered harmonics are shown separately in the two additional spectrograms. Interleaving of the odd and even harmonics (shown in the bottom spectrogram) reduces the rate of all-harmonic glissandi to one-half the rate heard with either set of components alone (and also drops the lowest pitch of the glides by 1 oct). For further details, see text.

third quarters should have superimposed rippling associated with cophasic and antiphasic repetition pitches, each with the same value of τ. The envelopes of the rippled power spectra of cophasic and antiphasic repetition pitches are complementary, with peaks of one coinciding with troughs of the other. When combined during all-harmonic glissandi, these antiphasic and cophasic ripples cancel,[4] and the only remaining spectral rippling corresponds to the separation from alignment of the two all-harmonic waveforms. Thus the all-harmonic repetition pitch falls until the end of the second quarter and then rises until the end of the fourth quarter of the cycle.

VIII. MISCELLANEOUS OBSERVATIONS

A. Perception of complex patterns produced by uncorrelated tone pairs

Uncorrelated complex tones (i.e., tones having independently assigned random phases for corresponding harmonic components) cannot produce pitch glides when mixed. There can be no movement toward or away from alignment of corresponding portions of the waveforms (and no corresponding systematic spectral interaction) as with mistuned correlated tone pairs (see Fig. 2). However, mistuned complex tones do produce patterns of amplitude fluctuation which are heard to repeat at a rate equal to the difference in fundamental frequency of the tones.

When 198- and 200-Hz uncorrelated complex tones were mixed, a complex amplitude pattern was heard clearly to repeat at 2 Hz. The beating of the fundamentals of the complex tones at 2 Hz was not necessary for hearing this repetition frequency, since a pattern was perceived to repeat twice a second with undiminished clarity when the fundamentals were absent. There appear to be two different mechanisms responsible for the 2-Hz iteration, one operating with unresolved upper harmonics, the other with resolved lower harmonics.

When the 198- and 200-Hz uncorrelated complex tones lacked the first nine harmonics (they were synthesized from polynomial equations and consisted of harmonics 10–40, each at equal amplitude), a clear 2-Hz periodicity was heard. When a tunable 1/3-oct filter (approximating a critical band) was swept slowly through the spectrum of the mixed complex tones, the 2-Hz repetition could be heard at all center frequencies. Examination of the waveforms of these bands showed periodic complex envelopes which were repeated each 500 ms at all center frequencies.

When the synthesized uncorrelated 198- and 200-Hz tones each consisted only of the first seven harmonics, a complex 2-Hz pattern was again heard clearly. In addition, some (but not all) of the harmonically related simple beat rates of the resolved lower harmonic pairs could also be heard. When listeners could hear the beating of individual harmonic pairs, they were generally modulated or accented at the 2-Hz frequency. As with the mistuned pairs of broadband and high-pass complex tones, the ensemble complex beat rate of 2 Hz did not require the presence of the simple 2-Hz beat rate of the spectral fundamentals, for a pattern repeating twice a second was still heard clearly when only harmonics 2–7 were present. In order to minimize the possibility that the 2-Hz periodicity resulted from the interaction of harmonic distortion products, listening was carried out at low levels. When the signal was 15 dB SL, the 2-Hz periodicity was heard clearly by all listeners. An 80- to 300-Hz bandpassed white noise (filter slopes 96 dB/oct) was then introduced at 15 dB SL and the signal readjusted to 15 dB SL in the presence of the noise (which would be expected to mask any low level distortion products). Once again, all listeners heard a 2-Hz repetition matching the frequency of the missing fundamental beat rate (the beat rates present in the acoustic signal were 4, 6, 8,..., 14 Hz). It should be emphasized that the ensemble complex beat produced by the lower harmonics was quite different from the simple waxing and waning of amplitude produced by a pair of beating sinusoidal tones, consisting rather of a complex periodic pattern of amplitude modulations. Unlike the 2-Hz iterance described earlier for 1/3-oct bands of unresolved harmonics, the 2-Hz periodicity heard with only the first seven harmonics involved the integration of harmonically related periodic patterns across different neural frequency channels.

B. Glissandi involving vowels and other special sounds

In keeping with the concept that iterated randomly-derived waveforms can serve as exemplars or model periodic stimuli, the observations reported can be applied to other types of periodic stimuli as well (keeping in mind any special characteristics of particular sounds). Correlated pairs of broadband complex tones produced by standard laboratory generators (pulse trains, sawtooth waves, etc.) can be used to produce pitch glides. However, correlated tones consisting of only odd harmonics (e.g., square waves) produce glissando pairs (rising and falling glides) at rates twice that of all-harmonic tones (e.g., unipolar pulse trains) having the same extent of mistuning. Glissandi can also be heard for mistuned vowels (although the glides are somewhat weaker than for broadband stimuli lacking pronounced formants): When a single glottal pulse of the vowel "ee" was repeated on two digital delay lines, and the clock frequency driving one delay line was changed slightly, glissandi were heard by all of our listeners. Faint pitch glides were heard when the vowel was mistuned slightly from unison with a pulse train (all harmonics in cosine phase). Apparently, glottal buzzes, even after passage through the vocal tract, are sufficiently similar to pulse trains to cause listeners to hear systematic pitch changes. However, any particular broadband complex tone (whether pulse train, vowel, or other) when mixed with an iterated frozen noise segment formed an uncorrelated pair, and no hint of glissandi could be heard.

ACKNOWLEDGMENTS

The support of grants from the Air Force Office of Scientific Research and the National Institute of Mental Health is gratefully acknowledged. We thank Keri Riener, Jill Robertson, and David Taffet for their valuable contributions.

[1] In an earlier study, we reported that several other perceptual phenomena which have been reported for periodic pulse trains (and attributed to their pulsate nature) can also be observed with periodic nonpulsate sounds (Warren and Wrightson, 1981).

[2] It has been suggested that the tonal and infratonal repetition together form a single perceptual continuum of detectable acoustic repetition called "iterance" which extends from a lower limit of 1 Hz through the upper limit of audibility at about 16 000 Hz. There is evidence indicating that mechanisms subserving iterance have some degree of overlap in the tonal and infratonal ranges (see Warren, 1982, pp. 80–85).

[3] While observing the spectral changes on a real-time spectrum analyzer, pitch glides corresponding to changing spectral maxima of harmonics, or groups of harmonics were heard with careful listening. These glides were more evident at high sensation levels, covered short frequency ranges, and were much briefer than the major glissandi.

[4] Spectral analysis shows that the rippling of the power spectrum envelopes corresponding to antiphasic and cophasic repetition pitch can cancel when superimposed. We used two uncorrelated broadband Gaussian noises to demonstrate this cancellation. Each of the noises was delayed by the same period τ and then mixed with itself, one addition being cophasic and the other antiphasic. Characteristic antiphasic and cophasic repetition pitches were heard for each when presented separately. But when these two comb-filtered noises were mixed at equal amplitudes, the ripples canceled, a flat spectrum was produced, and no repetition pitch could be heard.

Bilsen, F. A. (1966). "Repetition pitch: Monaural interaction of a sound with the repetition of the same, but phase shifted, sound," Acustica **17**, 295–300.

Hartmann, W. M. (1985). "The frequency-domain grating," J. Acoust. Soc. Am. **78**, 1421–1425.

Nábělek, I.V., Nábělek, A. K., and Hirsh, I. J. (1970). "Pitch of tone bursts of changing frequency," J. Acoust. Soc. Am. **48**, 536–553.

Nábělek, I.V., Nábělek, A. K., and Hirsh, I. J. (1973). "Pitch of sound bursts with continuous or discontinuous change of frequency," J. Acoust. Soc. Am. **53**, 1305–1312.

Plomp, R. (1964). "The ear as a frequency analyzer," J. Acoust. Soc. Am. **36**, 1628–1636.

Plomp, R. (1976). *Aspects of Tone Sensation* (Academic, London).

Plomp, R., and Mimpen, A. M. (1968). "The ear as a frequency analyzer. II," J. Acoust. Soc. Am. **43**, 764–767.

Small, A. M., Jr., and McClellan, M. E. (1963). "Pitch associated with time delay between two pulse trains," J. Acoust. Soc. Am. **35**, 1246–1255.

Thurlow, W. R., and Small, A. M., Jr. (1955). "Pitch perception for certain periodic auditory stimuli," J. Acoust. Soc. Am. **27**, 132–137.

Trenque, P., and König, E. (1988). "The chirp train streaming test: A clinical approach to auditory selectivity and listening," Audiology **27**, 65–88.

Warren, R. M. (1978). "Complex beats," J. Acoust. Soc. Am. Suppl. **64**, S38.

Warren, R. M. (1982). *Auditory Perception: A New Synthesis* (Pergamon, Elmsford, New York).

Warren, R. M., and Bashford, J. A., Jr. (1988). "Broadband repetition pitch: Spectral dominance or pitch averaging? " J. Acoust. Soc. Am. **84**, 2058–2062.

Warren, R. M., Brubaker, B. S., and Gardner, D. A. (1984). "Perception of complex tone pairs mistuned from unison: Waveform relations determine whether pitch glides or iterated complex auditory patterns are heard," J. Acoust. Soc. Am. Suppl. 1 **75**, S20.

Warren, R. M., and Wrightson, J. M. (1981). "Stimuli producing conflicting temporal and spectral cues to frequency," J. Acoust. Soc. Am. **70**, 1020–1024.

Wilson, J. P. (1966). "Psychoacoustics of obstacle detection using ambient or self-generated noise," in *Animal Sonar Systems*, edited by R. G. Busnel (Louis-Jean, Gap, Hautes-Alpes, France), pp. 89–114.

Yost, W. A., Hill, R., and Perez-Falcon, T. (1978). "Pitch and pitch discrimination of broadband signals with rippled power spectra," J. Acoust. Soc. Am. **63**, 1166–1173.

Appendix F

THE FIRST INTERNATIONAL CONFERENCE ON
MUSIC PERCEPTION AND COGNITION
KYOTO, JAPAN 17-19 October 1989

MELODIC AND NONMELODIC PITCH-PATTERNS: EFFECTS OF DURATION ON PERCEPTION

Richard M. Warren, Daniel A. Gardner, Bradley S. Brubaker & James A. Bashford, Jr.

Department of Psychology
University of Wisconsin-Milwaukee
Milwaukee, Wisconsin 53201, U.S.A.

Notes used for melodic themes have durations extending roughly from 150 to 900 ms (Fraisse, 1963, p. 89). Experiment 1 examined the effects of going beyond these limits on melodic perception, employing notes from 40 ms to 3.6 s. Experiment 2 examined the perception of simple nonmelodic sequences using tones with durations from 10 ms through 5 s.

Our first experiment employed phrases consisting of the first 7 to 9 notes of a familiar melody. Durational limits for recognition were determined for listeners (30 subjects without formal musical training recruited from introductory psychology classes) who tried to identify each of eight melodies. Before starting the experiment, subjects were screened to determine if they could name the melodies presented at moderate tempos (320 ms/note) and repeated several times without pauses. Listeners who could not identify at least 6 of the 8 melodies were rejected. Those who passed this screening served in the formal experiment. Each melody was presented in a series of increasing note durations starting with 40 ms/note. Steps were increased by √2 (successive increases of 41%) until recognition was accomplished, or the limit of 320 ms/note was reached. The same subjects also received melodies with long-duration notes, starting at 3.6 s/note with successive presentations involving steps decreasing by a factor of the √2, again until recognition was accomplished or durations reached 320 ms/note. Values cited for notes always included 20 ms of silence. Half the subjects started with the series of increasing item durations, and half started with the series of decreasing item durations. In preliminary experiments, we found that when durations were greater than two s/note, although listeners could not identify melodic patterns directly, they sometimes could guess correctly by recognizing one or two of the intervals formed by contiguous tones. This inferential recognition worked best when based upon the initial notes of the melody, much as it is easier to think of a word if we are given the first two or three letters rather than the same number of letters found elsewhere in the word. In order to make this type of guessing more difficult, the repeated melodies were started with either the third or fourth note.

The results are shown in Figure 1. It can be seen that the grand medians for the lower and for the upper limits for melodic recognition (160 ms and 1280 ms respectively) approximate the limits of the range of durations customarily employed for melodies (from roughly 150 to 900 ms) as cited by Fraisse (1963).

Do the lower limits for recognition found in Experiment 1 represent a general inability to recognize particular sequences of tones when their durations are very brief, or do they reflect a special limit for melodic perception? Experiment 2 departed from the use of melodies, and determined whether a separate group of 36 subjects (again, students from introductory psychology courses) could distinguish between sequences of three tones arranged in different orders. Since special cues to order are provided by the ease of identifying the first and last items of a sequence (Warren, 1972), these cues to temporal arrangement were avoided by iterating the sequences. The simplest repeated sequence for which the same items can be arranged in different ways consists of three elements (A, B, and C) which can be arranged as ABCABCABCA... or ACBACBACBA... . The three tones chosen for our study were sinusoids having frequencies of 988, 1661, and 2791 Hz. Neighboring frequencies differed by 9 semitones, a separation which is considerably greater than the value of approximately 2 or 3 semitones cited by Dowling and Harwood (1986, p. 83) as the width of critical bands in this frequency range. Hence, it may be assumed that separate populations of receptors were stimulated by each tone. Listeners heard a succession of sequence bursts which either were all identical [ABCABC...A, pause, ABCABC...A, pause, etc.] or which alternated between the two possible arrangements [ABCABC...A, pause, ACBACB...A, pause, ABCABC...A, pause, etc.]. They were required to tell whether successive bursts were the same or different. Presentation was always in order of increasing item duration as shown in Table 1, which also lists the stimulus parameters.

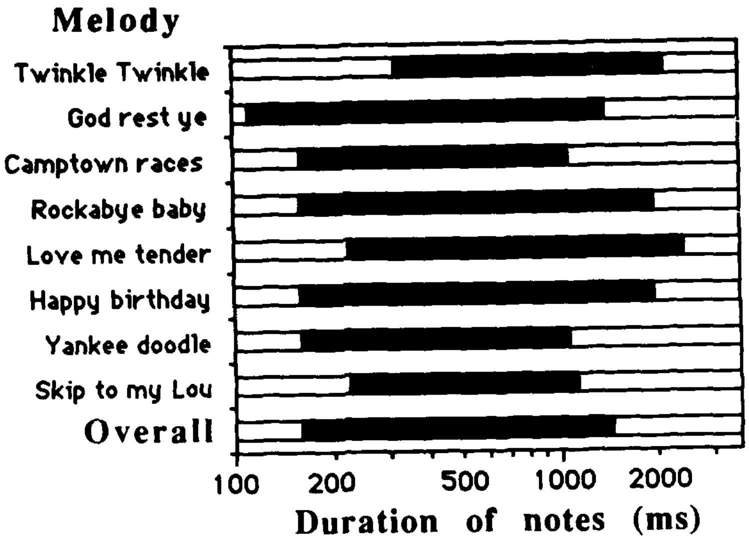

Figure 1. Note durations required for melodic perception. The limits of this range for eight melodies, shown by the shaded bars, represent medians based upon judgments of 30 subjects without formal musical training. Below this limit, while distinctive patterns are heard for each of the sequences of notes, the corresponding melodies cannot be perceived. Above this limit, the long notes are heard only as sequences of individual pitches.

Subjects listened for as long as they wished to the successive bursts before judging whether alternate bursts were identical or different in any way. Three same/different judgments were made by each of the 36 subjects at each stimulus duration. The experimenter delivered the same and different stimuli in a balanced pseudo-random order.

The results obtained are shown in Table 2. It can be seen that performance was well above chance at all item durations.

Table 1. Sequence Parameters in Experiment 2

Item	Items per	Inter-burst	Bursts	Burst
Duration	Series	Interval	per	Duration
(ms)	Burst	(ms)	Series	(sec)
10	301	300	10	3.01
30	91	300	10	2.73
100	31	300	8	3.10
300	10	300	8	3.00
1000	10	1000	4	10.00
3000	7	1000	4	21.00
5000	7	1000	4	35.00

Table 2. Accuracy of Same/Different Judgements in Experiment 2

Duration	Number	Percent	Z Score
of Item	Correct	Correct	
(ms)	(out of 108)		
10	93	86	7.51*
30	93	86	7.51*
100	100	93	8.86*
300	99	92	8.66*
1000	103	95	9.43*
3000	98	91	8.47*
5000	103	95	9.43*

*Significant at $p < .0001$

The results obtained in Experiment 1, together with other evidence, suggest that melodies can be perceived only if they fall within the limits of a familiar "temporal template." The bounding durations of the template used for recognition are characteristic of melodic sequences, and are not characteristic of tonal sequences in general. Thus, in Experiment 2 we found that discriminating between different arrangements of the same tones in simple nonmelodic sequences could be accomplished readily even when the interval separating the onset of successive tones was as brief as 10 ms or as long as 5 s.

Below 100 ms, it was impossible to perceive the order of items in Experiment 2, and judgments were made on the basis of differences in "quality." This is in keeping with informal observations involving the stimuli used in Experiment 1. It was noted that the sequences of tones presented at durations below the threshold for melodic identification each had distinctive characteristics unrelated to the melodies emerging at longer durations. It would appear that the lower limit for melodic identification is not due to the "metathesis" of notes which then results in a perceptual confusion of different orders as suggested by Winckel (1967).

Why was performance excellent at five second item durations in Experiment 2, while melodic identification in Experiment 1 had an upper limit well below this value? In Experiment 2, each of the six possible pitch transitions (e.g., 2791 Hz to 988 Hz) was present in only one of the arrangements of the 3-item repeated sequences. Remembering any one of these pitch changes heard with long duration tones would suffice for determining whether the sequences were the same or different. In Experiment 1, the task of melodic identification when the notes lasted a few seconds was quite different. Melodic cohesion was lost, and each note was heard as an independent sound. Reconstruction of the melodies from these successive tones was difficult without formal training in music. One method reported by a few listeners involved remembering the intervals from one tone to the next as they occurred at the slow rates, and then "playing back" from memory at a faster rate. People with formal training used musical notations as an aid, and could reconstruct a melody readily, much as a word can be identified by literate listeners when they hear it spelled.

Acknowledgments. This study was supported by grants from the National Institutes of Health and the Air Force Office of Scientific Research.

References

Dowling, W.J., and Harwood, D.L. Music cognition. (Academic Press, New York, 1986).

Fraisse, P. The psychology of time. (J. Leith, Trans.). (Harper & Row, New York, 1963).

Warren, R.M. "Perception of temporal order: Special rules for initial and terminal sounds of sequence". Journal of the Acoustical Society of America, 52, 167, 1972 (Abstract).

Winckel, F. Music, sound and sensation: A modern exposition. (Dover, New York, 1967).

Appendix G (in press, Perception & Psychophysics)

TWEAKING THE LEXICON: ORGANIZATION OF VOWEL SEQUENCES INTO WORDS

Richard M. Warren

James A. Bashford, Jr.

Daniel A. Gardner

University of Wisconsin-Milwaukee

Department of Psychology

Milwaukee, Wisconsin 53201

Mailing Address: c/o R.M. Warren
Department of Psychology
University of Wisconsin-Milwaukee
Milwaukee, Wisconsin 53201

Telephone No. (414) 229-5328

Running Head: Tweaking the Lexicon

Abstract

The ability of listeners to distinguish between different arrangements of the same three vowels was investigated for repeating sequences having item durations ranging from 10 ms (single glottal pulses) up to several seconds/vowel. Discrimination was accomplished with ease by untrained subjects at all item durations. From 30 ms through 100 ms/vowel, an especially interesting phenomenon was encountered--the sequences of steady-state vowels were organized into words, with different words heard for the different arrangements of items. A second experiment employed repeating sequences consisting of random arrangements of ten 40-ms vowels. When sets of four such sequences were presented to listeners, distinctive words were heard which permitted each arrangement to be discriminated from the others. Further, minimal differences (reversing the order of a single contiguous pair of vowels) in the ten-item sequences could be detected via verbal mediation. Hypotheses are offered concerning mechanisms responsible for these results.

Tweaking the Lexicon: Organization of Vowel Sequences into Words

Introduction

A succession of steady-state vowels presented loudly and clearly can be heard as a word. This unusual verbal organization can help us understand how acoustic components are processed in speech perception.

The experiments reported here employed continuously repeated or recycled sequences. These iterated stir uli were first used in the 1960's as a means of allowing a limited number of sounds (usually 3 or 4) to be presented for extended periods (Warren, 1968; Warren, Obusek, Farmer, & Warren, 1969). Repetition also helps to avoid the special cues to the order of items that are provided by the first and last items of a sequence (Divenyi & Hirsh, 1978; Warren, 1972). Studies employing recycled vowel sequences have shown that the naming or identification of order is accomplished readily at 200 ms per item, but is not possible at item durations below 100 ms (Cole & Scott, 1973; Cullinan, Erdos, Schaefer, & Tekieli, 1977; Dorman, Cutting, & Raphael, 1975; Thomas, Cetti, & Chase, 1971; Thomas, Hill, Carroll, & Garcia, 1970; Warren, 1968; Warren, Obusek, Farmer & Warren, 1969; Warren & Warren, 1970). None of these studies reported observations involving vowel durations below the threshold for identification of order. However, in preliminary observations we found that when three steady state vowels (A, B and C) were presented as recycled sequences, the two possible arrangements (...ABCABCA... and ...ACBACBA...) could be discriminated readily at item durations much briefer than the limit for naming of order.

Our first experiment confirmed that discriminating between different orders of the same speech sounds does not require the ability to identify the order of the phonemes, or indeed even the ability to identify the components within the sequences. Listeners were required to judge whether alternately presented recycled sequences of three vowels (which could be presented in identical or permuted item orders) were "same" or "different." The vowels spanned the range from 10 ms (single glottal pulses) through 5 s (500 glottal pulses), with no acoustic mixing or transitional stages in going from one vowel to the next. When vowel durations were above 100 ms, listeners could name the phonemes in the proper order, and then use the difference in named order to distinguish the two arrangements. When vowel durations were below 30 ms, resemblance to speech was absent, and differences in quality or timbre made it easy to discriminate between the two arrangements (for example, a listener might report that one sequence sounded "rougher" than the other). Between these values (from 30 ms through 100 ms), listeners could hear different words (usually lexical, sometimes nonsense) for each of the arrangements. The words heard differed across individuals, and normally bore little resemblance to the actual phonemes.

Other studies also have observed verbal organization of repeated sequences of vowels. Dorman et al. (1975) used these stimuli to investigate the limits for identification of temporal order and noted in passing that verbal organizations interfered with the listeners' task when vowel durations approached the lower limit for order identification. Skinner (1936) used repeated sequences of barely audible vowels having durations of several hundred milliseconds. He reported that his listeners heard words and sentences. When the level of vowels were raised well above threshold (as were the vowels in each of our experiments), "imitative" responses occurred, and the sequences were identified as a succession of vowels (in keeping with observations made in our experiment 1 for vowels having durations of a few hundred milliseconds). Skinner attributed the verbal organization of his faint syllabic-length vowels to a "summation" of originally subliminal responses.

Our second experiment employed 48 recycled sequences, each consisting of a different random arrangement of a set of ten steady-state 40-ms vowels played loudly and clearly. Individual listeners heard characteristic words or pseudowords corresponding to each of the orders, and could identify a particular sequence among several on the basis of its verbal correlate. Interestingly, listeners often heard a particular arrangement as two concurrent words that differed in timbre and/or pitch. As we shall see, this splitting of the stimulus provides a clue to the mechanisms employed for perceptual syntheses.

Another part of the second experiment also employed recycled sequences of ten different 40-ms vowels. The stimuli consisted of pairs of sequences having minimal differences in structure (the order of two contiguous vowels was interchanged). Listeners were again able to use verbal mediation to distinguish members of the pairs.

Experiment 1: Discriminating between Different Orders of Three-Item Vowel Sequences

This first experiment was designed in part to test the hypothesis that discriminating between different orders of the same speech sounds does not require the ability to identify the order of the phonemes, or indeed even the ability to identify the components within the sequences. Our listeners were required to judge whether alternately presented recycled sequences of three vowels (which could be presented in identical or permuted item orders) were "same" or "different." The vowel durations extended from 10 ms to 5 s, permitting a comparison of discrimination accuracy and cognitive strategies employed for durations corresponding to, as well as briefer and longer than, those occurring in speech.

Method

Subjects. Participants were recruited from introductory psychology courses, and received either course credit or cash for their participation. Students who passed the audiometric screening procedure described below were assigned randomly to one of two experimental groups, each containing 36 subjects.

Audiometric Screening. All subjects participating in the experiments passed an audiometric test designed to eliminate individuals with hearing deficits, as well as those who failed to follow the standard instructions used with the Békésy threshold tracking procedure. Following presentation of instructions and familiarization with the task, a pure tone presented diotically was swept up from 400 Hz through 9000 Hz and then down from 9000 Hz through 400 Hz at a rate of one octave/minute while subjects tracked their thresholds. Tracking was accomplished using a hand-held button switch (depressing the button decreased the intensity at a rate of 2.5 dB/s and releasing the button increased the intensity at the same rate). An X-Y plotter produced audiograms consisting of continuous threshold tracings. Subjects were excluded from further participation if either directional sweep resulted in audiograms which differed from the 1964 ISO standards by more than 22.5 dB at any frequency for the portion of the audiogram extending from 500 through 8000 Hz.

Stimuli. The first stage in the preparation of the recycled sequences of three vowels used as stimuli involved production of extended steady-state recordings of three vowels (/ʌ/, /æ/, /i/) on parallel tracks of a multitrack recorder (16-track Ampex model MM 1200). These steady-state vowels were derived from recorded statements of syllables containing these vowels ("hud" for /ʌ/, "had" for /æ/, and "heed" for /i/) produced by a male speaker at a vowel fundamental frequency of 120 Hz (the speaker matched the pitch of the vowel to that of a complex tone of 120 Hz heard through headphones). A complete single glottal pulse was excised from the central portion of each consonant-vowel-consonant statement. The waveforms of the glottal pulses were monitored and the period measured by a Nicolet model 3091 digital storage oscilloscope used in conjunction with a programmable digital delay line (modified Eventide model BD955) capable of repeating or "looping" stored input corresponding to a single glottal pulse. The repetition period of the delay line was set at 8.33 ms for each of the vowels, (which corresponded to a repetition rate of 120 Hz), and recordings of the steady-state vowels were made on parallel tracks.

Two types of series were recorded for each duration employed (10, 12, 30, 100, 300, 1000, 3000, and 5000 ms). A "different" series with successive sequence bursts consisting of /ʌ/, /æ/, /i/, /ʌ/, /æ/, /i/, ... /ʌ/ alternating with the permuted order /ʌ/, /i/, /æ/, /ʌ/, /i/, /æ/, . . . /ʌ/, and a "same" series with all bursts consisting of /ʌ/, /æ/, /i/, /ʌ/, /æ/, /i/, . . . /ʌ/. Note that because of the special ease of identifying the first and last items of a sequence (Warren, 1972), each sequence began and ended with the same item. Table 1 lists the parameters for the stimuli employed, giving item durations, the number of items (vowel statements) in each sequence burst, the interburst interval separating successive bursts (which were either identical or alternating in item order), and the number of bursts constituting a stimulus set.

All sequences (except those with a 12 ms item duration) were generated by gating the output from the three parallel tracks containing extended steady-state single vowels prepared as described above. The output of these tracks was passed through three Coulbourn electronic switches set for a rise/fall time of 2 ms. A series of timers (Grason-Stadler model 1216A) controlled passage of each vowel through its gate, and introduced a 1 ms separation between the waveforms corresponding to the ending of one vowel and the beginning of the next as seen on the digital storage oscilloscope (this separation minimized acoustic interaction of items). Another timer regulated the silent interburst interval separating successive bursts. The path of the signals through the equipment was identical in both the "same" and "different" order series, with the relative timing of the opening and closing of the gates producing the permuted orders. The number of vowel statements in the burst was controlled by Coulbourn predetermined counters (model S43-30). The outputs of the gates were combined in a Yamaha audio mixer (model PM-430) and band-passed from 100 through 8000 Hz by a Wavetek/Rockland filter (model 751A) having slopes of 115 dB/octave before recording on one of the tracks of the multitrack recorder reserved for the experimental stimuli. The input level of each of the vowels as delivered to the tape recorder was adjusted separately to produce equal intensity (dBA) for all vowels on subsequent playback through the headphones used by the subjects. Sets of sequences were recorded for each item duration with the track for a "same" set parallel to the track of the corresponding "different" set, so that the experimenter could present either "same" or "different" stimuli at the same tape positions.

Insert Table 1 about here

In sequences consisting of 10 ms items, only a single statement of each vowel's waveform was gated before switching to the next. Since this programmed switching was not exactly synchronous with the waveform repetition period of the recorded vowels (the recorder had a frequency stability of ±0.1%), the repeated sequences underwent slow drifts in their waveforms and perceptual qualities. Sequences with longer item durations consisted of multiple identical statements of each vowel's waveform before switching to the next, and perceptual quality was more stable.

Sequences with 12 ms item durations were constructed with "locked" waveforms, so that switching always occurred at a fixed position in the waveform of each vowel and drifting did not take place. This stimulus was prepared as follows: Three separate delay lines (two modified Eventide model BD955's and one modified Eventide model 1745M) were driven by a common clock, and each repeated a single glottal pulse of a different vowel (the glottal pulses were derived from the same extended statements of the three vowels used for preparing the other sequences). The manner of capturing and repeating single glottal pulses on the delay lines was similar to that already described, except that the programming equipment introduced a 3.67 ms silent interval between successive statements of glottal pulses. The splice point of each single-vowel digital loop was at the center of this silent interval. The repetition period of all delay lines was set at 12 ms (measured by a common clock), the vowel statements were aligned so that each of the three vowels began and ended synchronously, and the vowels were then recorded simultaneously on separate tracks of the multitrack recorder. A timing signal (a unipolar pulse) generated by one of the delay lines at the splice point of its digital loop was recorded on a fourth track at the same time as the vowels. This recorded timing signal permitted the programming equipment to gate single glottal pulses of each recorded vowel in the desired order. Following gating and mixing, the locked three-item vowel sequences were recorded on an additional channel of the multitrack recorder, as described for the other sequences.

Experimental Procedure. Subjects passing the audiometric screening test were recalled for their single experimental session lasting about 40 minutes. They were tested individually while seated in an audiometric room along with the experimenter. Stimuli were presented diotically through matched headphones at a level of 70 dBA as measured by a sound level meter with a 6 cc coupler. The experimenter operated the tape recorder (located outside the chamber) using a remote control unit equipped with a preset multipoint rapid search-to-cue device. Switches on an audio mixer permitted delivery of the output from the desired track of the recorder.

Half of the 72 subjects served in the main experiment, which employed all the sequence pairs listed in Table 1 except for the sequences with 10 ms vowel durations (that is, 12, 30, 100, 300, 1000, 3000, 5000 ms item durations) presented in the order listed. The order of increasing item durations was employed to avoid the possibility (discussed by Warren, 1974a) that with a series of decreasing item durations, the naming of order at brief item durations could be accomplished through recognition of qualitative similarities to the previous sequences having longer item durations with directly identifiable orders. As described earlier, the 12 ms sequences had switching from one vowel to the next vowel "locked", so that each restatement of a particular vowel was a single intact glottal pulse. All other sequences were "nonlocked," with successive statements of each vowel starting and stopping at different waveform positions. The separate group of 36 subjects serving in the supplemental experiment received only the stimulus consisting of 10 ms vowels.

Subjects were told that they would be hearing patterns of sounds separated by brief silent intervals, and that their task was to determine if all patterns were identical or if alternate patterns differed in any way. They were instructed to call out "same" or "different" at any time during the stimulus presentation. They were informed that the occurrence of same and different groupings would be randomly determined. Subjects were encouraged to ask questions if any part of the instructions was unclear. After both the subject and experimenter were satisfied that the instructions were understood, the sequences were presented in order of increasing item duration for the 36 subjects in the main experiment (the 36 subjects serving in the supplemental experiment received only the 10 ms items).

Before presentation of unknown same or different sequences at each item duration, each subject was given sample sequences (first a "different" set, then a "same" set) which were identified by the experimenter as same or different. They were told that they could hear either of the known samples again, if they wished, before hearing the unknowns. When a subject indicated readiness, s/he was given three unknowns at that item duration. The same and different unknowns were presented in a pseudorandom order with the constraint that all three of the unknowns presented to a subject at any item duration could not all be of a single type. No feedback was given concerning the accuracy of judgments with the unknowns. In the main experiment, of the total of 21 unknowns presented to each subject, 10 were the same and 11 were different for 18 subjects, and 11 were the same and 10 were different for the other 18 subjects. Half the subjects received orders of same and different unknowns which were "mirror images" of the other half, with same and different unknowns being interchanged to maintain symmetry of unknown groupings. This symmetry was also maintained for the supplemental group receiving the 10 ms vowel durations.

Results

Table 2 shows that the accuracy of discriminating between permuted orders ranged from 78% correct to 99% correct, and was significantly better than chance for all of the item durations used ($Z \geq 5.77$, $p < .0001$). The sequences consisting of 10 ms items (with

Insert Table 2 about here

slowly drifting waveforms and perceptual qualities) had 78% correct responses, while the 12 ms "locked" sequences (with switching occurring at fixed points corresponding to the beginning and end of the single glottal pulse representing each vowel) had 91% correct responses. This difference was significant ($Z \geq 3.70$, $p < .005$).

Questioning of listeners after completion of the formal experiment indicated that two basic ways of discriminating between the different arrangements of items were used: 1) naming of components in their proper order for vowel durations greater than 100 ms; and 2) a holistic recognition of patterns without the ability to identify the order of components (or even the components themselves) for vowel durations from 100 ms down to 10 ms. The range from 100 ms to 10 ms consisted of two regions: 2a) from 100 ms to 30 ms the sequences of 3 vowels could be heard as words rather than steady state vowels, with different words heard for the different arrangements; 2b) below 30 ms the vowel sequences were heard as nonlinguistic sounds, *with different qualities associated with the different* arrangements. These perceptual categories (1, 2a, 2b) reported by untrained listeners agreed with observations made by laboratory personnel.

Discussion

Limits for the Naming of Order

The earliest experiments with recycled sequences of sounds were concerned with thresholds for identifying the order of component items (Warren, 1968; Warren, Obusek, Farmer & Warren, 1969; Warren & Warren, 1970). When four 200 ms sounds were used, listeners instructed to name the order of items performed at chance level with unrelated sounds consisting of noises, tones, and buzzes, but they could accurately name the order of vowels. Subsequent studies established that the threshold for identifying the order of unrelated sounds is 300 ms or more (Warren & Obusek, 1972), while the threshold for correctly ordering the pitches associated with sequences of four sinusoidal tones is between 125 and 200 ms/item (Nickerson & Freeman 1974; Thomas & Fitzgibbons, 1971; Warren & Byrnes, 1975). The lowest thresholds for four item sequences (about 100 ms/item) were obtained with vowels (Dorman et al., 1975; Thomas et al., 1971).

There seems to be general agreement that vowel order can be named at briefer durations than is possible with other sounds. Why this difference? Using evidence from several sources, it was proposed by Warren (1974) [and also suggested independently by Teranishi (1977)] that the time required for verbal labeling or naming of components in extended sequences was the threshold-determining stage in the identification of order: Since vowels have a name which is the same as the sound itself, no recoding is necessary (naming order can be accomplished through a simple echoic restatement of the stimulus items), and the time required for identifying order is minimal. Nevertheless, as discussed below, the threshold value of 100 ms seems too high for agreement with models considering that identification of phonemic order is necessary for the comprehension of speech.

Normal conversation has an average duration of speech sounds of about 80-100 ms; this duration drops to about 70 ms for oral reading, and some comprehension of "compressed speech" is possible at average phonetic durations of only 30 ms (for a brief summary of this literature, see Warren, 1982, pp. 119-120). Recognizing that there was a discrepancy between the rate of phoneme occurrence within intelligible speech and the ability to perceive order in a sequence of independently generated speech sounds, Wickelgren (1969) suggested that context-sensitive allophones facilitated temporal ordering. Coarticulation is, at least in part, an acoustic consequence of inertial and neuromuscular constraints upon the movement of the tongue and other articulatory organs from one position to the next, and Wickelgren considered that by recognizing particular allophonic forms it might be possible to identify more than one phoneme from a single speech sound. Thus, order could be identified at much briefer durations than would be possible for a succession of independent sounds. A number of experiments have demonstrated thrt coarticulation (and other cues increasing the resemblance of phonetic sequences to normal speech) does indeed facilitate the task of naming components and their orders (Cole & Scott, 1973; Cullinan et al., 1977; Dorman et al., 1975; Warren, 1968; Warren & Warren 1970). But in no case, even with coarticulation cues, could orders be identified at item durations below 100 ms. However, listeners can comprehend speech consisting of phonemes with average durations considerably less than 100 ms. One explanation for this discrepancy is that phonetic order is determined at some early level of linguistic processing which is not accessible for the naming of this order. Another hypothesis (which we favor) is that a determination of the order of component speech sounds is not necessary at any level of analysis for the recognition of words or for the comprehension of discourse. It is suggested that acoustic sequences need not function as perceptual sequences (that is, a succession of discrete sounds). Patterns formed by particular arrangements of speech sounds may be recognized as "temporal compounds" without the need for identification of constituents. As discussed below, this concept of temporal compound formation was formulated initially on the basis of experiments with nonverbal sounds.

Nonphonetic Temporal Compounds

In earlier studies involving arbitrarily selected sounds (noises, sinusoidal tones, and complex tones), listeners attempted to distinguish between different arrangements of repeated sequences consisting of the same three items which were presented without any acoustic interactions or transitions involving contiguous sounds (Warren, 1974; Warren & Ackroff, 1976). These studies demonstrated that the different arrangements of nonverbal sounds could be discriminated with ease for item durations from 5 through 100 ms--yet the naming of order was not possible within this range. It was suggested that permuted orders of brief items could be distinguished through the bonding of components to form "temporal compounds" possessing characteristic qualities, even though the component acoustic elements and their arrangements could not be identified. Thus, for an isomeric pair of temporal compounds consisting of identical components arranged in different orders, a listener might describe one arrangement of the nonverbal sounds as "bubbly" and the other as "shrill." These qualitative differences served as the basis for accurately differentiating between different acoustic orders.

Vowel Sequences and Their Verbal Temporal Compounds

Our subjects in experiment 1 indicated that discrimination of permuted orders was accomplished in different ways at different item durations. When the item durations corresponded to single glottal pulses (10 and 12 ms vowels), listeners used nonverbal temporal compounds to distinguish between the permuted vowels. Thus, with these very brief durations an individual might report, for example, that one order was characterized by a "dull" quality while the other order sounded "crisp." However, perceptual organization into syllables and words (verbal temporal compounds) occurred at vowel durations from 30 through 100 ms. Within this durational range, a listener might say that one arrangement of vowels resembled or brought to mind repetitions of the word "kettle" while the other arrangement sounded more like repetitions of "puddle"--this despite the great phonetic differences between the actual stimuli and their lexical correlates. The specific word corresponding to a particular temporal arrangement varied from listener to listener.

It appeared desirable to study further the verbal organization of a succession of steady-state vowels into words, and experiment 2 was undertaken with this purpose.

Experiment 2a: Identifying Different Arrangements of Ten-Item Vowel Sequences

Experiment 1 has shown that recycled sequences of steady-state vowels played loudly and clearly can be heard as coherent verbal utterances, and that different arrangements of the same vowels can be discriminated on the basis of their distinctive verbal organizations. Further informal observations indicated that roughly 30-80 ms/vowel was the optimal duration for hearing words. Experiment 2a was designed to examine the characteristics of this vowel-word illusion using recycled sequences consisting of ten 40-ms vowels. The 400 ms duration of these sequences corresponded to that of words in normal conversation. During the experiment, listeners were presented with four recycled sequences each having a different randomly determined vowel order, and they were instructed to use verbal organizations as a means of identifying the different patterns on second presentation.

Method

Subjects. Thirty-two auditometrically screened listeners (14 male and 18 female) were recruited from introductory psychology courses, and received either course credit or cash for their participation. The screening procedure was the same as that described for experiment 1.

Stimuli. For synthesis of the 10 vowel components, a Data Precision model 6100 Universal Waveform Analyzer, operating at a sampling rate of 40 kHz with 14-bit resolution, was used to excise single 5 ms glottal pulses from a male speaker's sustained productions (200 Hz voicing frequency) of ten vowels (those in 'heed', 'hid', 'head', 'had', 'hod', 'hawd', 'hood', 'hud', 'hoot' and 'herd'). The digitized glottal pulses were then iterated 8 times to produce 40 ms bursts that were judged by a panel of four trained listeners to be identifiable as the parent vowels. Linear ramps of 2.5 ms (zero dB minimum) were imposed upon the onset and offset of each vowel burst for suppression of transients, and the amplitude envelopes of the bursts were rescaled so that each would play back at the same level.

The ten vowel bursts were sampled randomly without replacement and concatenated in digital form to create 48 ten-item sequences (out of a total of factorial 9 possible orderings). Digital-to-analog conversion and playback of the 400-ms sequences in recirculating form was accomplished using a Data Precision Co. Polynomial Waveform Synthesizer model 2020-100 (40kHz sampling frequency with 12 bit resolution). The analog playback of the recycling sequences was recorded on an Otari model MTR 90-II 16-track recorder, with sequences to be presented on the same trial (four sequences for each of the 12 trials) recorded in parallel on separate tracks. During the experiment, the output of the recorder was amplified by a Neotek Series I audio mixer and bandpass filtered from 50 Hz to 8000 Hz with slopes of 115 dB per octave (Wavetek/Rockland model 751A Brickwall Filter).

Experimental Procedure. Listeners were tested individually in an audiometric room with stimuli delivered at 70 dBA SPL through diotically wired TDH-49P headphones mounted in MX 41/AR cushions. The experimenter operated the Otari recorder (located outside the chamber) using a remote preset search-to-cue device. Switches on the audio mixer located inside the chamber permitted delivery of the output from the desired tracks of the recorder.

Listeners participated in two practice trials and ten formal trials, with the 12 sets of sequences presented in the same order to all listeners. Each trial consisted of a learning phase and a test phase. During the learning phase, listeners were presented successively with four sequences, and were required to listen to the recycling vowel patterns until they could write down what the voice seemed to be saying. (For their transcriptions, listeners used a response booklet having separate pages for each experimental trial.) It was explained that their written descriptions would provide a means of identifying the sequences during the test phase of the trial. Once the listener had provided written responses for each of the 4 sequences, s/he began the test phase using a control box having buttons labeled A, B, C, and D. Each of the buttons could be used to deliver one of the four sequences presented during the learning phase of the trial, and the listener's task was to match the letter of each button with their previous verbal organization for that sequence. Listeners did this by placing appropriately lettered cards beside their previous transcription.

Listeners were permitted to switch at will from one sequence to another during a trial's test phase, and they were given as much time as needed to complete the card-placing task. When matching was complete, the experimenter recorded the listener's response, provided feedback concerning accuracy, and began the next trial. During the debriefing period that followed the tenth formal trial, the experimenter reviewed the transcriptions to verify pronunciation and asked general questions concerning the listener's responses.

Results

Despite the obvious initial doubt of most listeners that they could accomplish the experimental task, perceptual organization of the recycling vowel sequences into syllables and words proved nearly effortless with little practice: The time required for initial verbal organization (that is, writing down a description for a particular vowel sequence) decreased from an average of about 35 s for the first practice sequence to an average of only 8 s per sequence across the ten formal trials. Further, once formed during the learning phase of a trial, these perceptual organizations proved sufficiently distinct and stable to permit rapid and highly accurate identification of the different vowel orderings during the test phase. On average, listeners completed the four matches of the test phase in about 15 s, and a majority of their responses were accurate even for the practice trials. Table 3 lists each trial separately, and gives the numbers of listeners who identified correctly each of the four sequences for the individual trials.

Insert Table 3 About Here

The chance likelihood of correctly identifying all four sequences on a trial was 1/24, so each fully correct series of responses by a listener exceeded chance at the .05 level. As can be seen, listeners identified all four sequences with above chance accuracy on most (better than 94%) of their attempts across the ten formal trials, with no evidence of fatigue or interference due to earlier sequences.

For the forty sequences presented in the formal trials, 35% of listeners' responses were nonlexical syllables (which always followed the rules for phoneme clustering of English), and the remaining 65% were words and phrases. Interestingly, most listeners also reported that certain sequences were organized as two different words (e.g., "Frankie" and "go animal") that sounded as though they were produced simultaneously by voices differing in quality. Despite the fact that the sequences were presented in the same order to all listeners, there was very little intersubject agreement in the forms reported for specific vowel orderings. Thus, although the verbal organizations were formed rapidly and were sufficiently stable to permit later recognition of sequences, they were also highly idiosyncratic - - perhaps due in part to the fact that the sequences were played as endless loops with no initial and terminal components.

Experiments 1 and 2a have shown that verbal mediation in the discrimination of random vowel sequences can be very robust when differences in order are substantial. Experiment 2b was designed to determine whether lexical matching could be extended to the discrimination of minimal differences in order.

Experiment 2b: Discrimination of Minimal Order Differences With Ten-Item Vowel Sequences

In the previous experiments, permuted orders of brief vowels produced distinct verbal organizations, but the differences in order were typically quite extensive: The two contrasting three-item sequences used in experiment 1 (ABCA ... and ACBA ...) had each of the three pairwise orderings of vowels reversed (AB vs. BA, BC vs. CB, and CA vs. AC), and in experiment 2, the 48 ten-item sequences were drawn without constraint from a pool of 362,880 possible recycled orders. In the present experiment, listeners made ABX judgments (deciding whether the unknown "X" was the same as "A" or "B"} for ten-item vowel sequences in which A and B differed only in the ordering of two contiguous vowels. Listeners also reported the basis for their discriminations for each trial.

Method

Subjects. Four subjects participated in the study. Subjects BB and JB were psychoacoustically trained listeners, and had participated in preliminary observations with ten-item sequences. Listeners JR and KR were not psychoacoustically trained and had no prior experience with the stimuli employed in this study.

Stimuli. Each of the 48 sequences used in experiment 2a was used as sequence "A" of a contrasting pair. The "B" sequence of each pair was produced by interchanging the order of two randomly selected contiguous vowels of the 10-item "A" sequence. Analog playback of the B sequences was recorded on the same 16-track tape used for experiment 2, with corresponding A and B sequences arranged in parallel. As in experiments 1 and 2, stimuli were amplified using an audio mixer and bandpass filtered from 50 Hz to 8000 Hz with slopes of 115 dB/oct.

Procedure. As in the earlier experiments, listeners were tested individually in an audiometric room with stimuli delivered through headphones at 70 dBA SPL. They were provided with a three-button panel which they used for switching between contrasting "A" and "B" stimuli and a third "X" stimulus which matched the sequence presented in either the A or B channel. Listeners switched at will between the three signals (each recorded on a separate track) until satisfied that they had determined which signal matched "X". After calling out either "A" or "B", they attempted to describe the basis for their discrimination. Listeners were aware that their ABX matches were being timed, and they received trial-by-trial feedback concerning their matching accuracy.

Listeners participated in a total of 16 sessions, with each session lasting about 20 minutes and involving judgments of six pairs of contrasting A and B sequences. Across the 16 sessions of each experiment, the forty-eight sequence pairs were presented twice to each listener for a total of 96 judgments. Each listener received a different random ordering of stimuli for their first ABX judgments of the sequence pairs, and this order was repeated for the listener upon second presentation of the stimuli, so that the two judgments for each contrast were separated by judgments of the remaining 47 sequence pairs.

Results

The number of correct responses (out of 96) and the median response times for judgments of each listener are presented in Table 4. As shown, overall matching accuracy was well above chance for all listeners, with the percentage of correct responses ranging from about 96% to 98%.

Insert Table 4 About Here

Listeners' trial-by-trial reports concerning the nature of their discriminations indicated that, although listeners attributed some discriminations to contrasting nonverbal characteristics (typically, differences in rhythmic complexity), most of the judgments were based upon differences in verbal organizations. These occasionally corresponded to pseudowords, but more often real words (e.g., "valuable" vs. "technical"). Most interestingly, although there was little agreement across listeners in the verbal forms evoked by specific vowel sequences, there was substantial consistency within listeners: in 52% of the cases in which listeners reported specific words upon first presentation of a contrasting pair of sequences, they reported the same word or words on second presentation of the sequences. This repetition of responses occurred in spite of the fact that successive judgments of the same stimuli were separated by several days and by interpolated judgments of the remaining 47 sequence pairs. Thus, although the verbal correlates of these monotone vowel patterns were again found to be highly idiosyncratic, they were also remarkably stable.

Discussion

Studies with ten-item sequences of nonverbal sounds have also found that minimal changes in ten-item sequences can be discriminated. Watson and his coworkers employed sequences of ten or more brief sinusoidal tones in experiments examining the ability to make fine discriminations (e.g., detecting a change in the frequency of a single tone) within complex "word-length" patterns (see Watson, 1987, for a review). These studies, which employed contrasting sequences presented as single statements, found that listeners usually required many hours of training before discrimination could be accomplished. However, Bashford & Warren (1988) reported that when sequences of ten tones are recycled, then the discrimination of fine changes is very much easier, and can be accomplished in less than one minute in a ABX discrimination task. They found performance with minimal changes (inverting the order of two of the ten tones) to be only slightly poorer than that observed with recycled ten vowel sequences. Hence, although perception in a "speech mode" (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967) is employed for sequences of vowels, fine discrimination can be accomplished with nonlinguistic sequences as well. Bashford and Warren also reported that sequences need not involve successions of discrete sounds for successful discrimination. Noise "sequences" were constructed by sampling from a catalogue of ten 40-ms segments which were previously excised from white noise. When the segments were abutted to form a loop, the recycled "sequence" resembled a repeated 400-ms segment of noise lacking the succession of discrete sounds characteristic of the sequences of tones and vowels. Interchanging the order of two contiguous 40-ms noise segments resulted in a discriminable change - - although ABX judgments took about twice as long as those with the recycled sequences consisting of ten 40-ms tones or vowels.

Let us look more closely at the linguistic organization occurring within sequences of vowels in Experiments 1 and 2. How is it that syllables and words are heard with a succession of steady-state vowels, despite the great differences between the phonetic compositions of the stimuli and the forms reported?

We hypothesize that the organization of our sequences of loud and clear vowels into syllables and words reflects shifts in perceptual criteria produced by repetition. The "criterion shift rule," which has been proposed for judgmental processing in general, considers that the criteria used for evaluating stimuli and events are displaced in the direction of simultaneous or recently experienced values (Warren, 1985). When applied to psycholinguistics, this effect can produce changes in the perceptual boundaries of phonemes following exposure to repeated syllables. While there is considerable controversy concerning the processes responsible for these boundary shifts (for discussion, see Diehl, Kluender, & Parker, 1985), there is agreement that the changes which do occur move the acoustic boundaries delineating particular phonemes toward a closer correspondence with the iterated stimulus. This shifting of criteria may be considerably greater when repetition is continuing (as in the present experiments) than after repetition ceases (as is typically the case in studies measuring the extent of category boundary shifts). It appears that in the present experiments the continuing repetition of a loud and clear sequence of steady-state vowels changed the acoustic requirements for recognition of a syllable or word to the point where the stimulus itself could be perceived as a particular utterance by a speaker.

The perceptual matching of a repeated vowel sequence to a particular verbal form may be facilitated not only by a criterion shift, but also by the splitting of the stimulus into two simultaneous percepts. Recall that typically an iterated vowel sequence splits into two concurrent forms - - usually two voices with different pitches or qualities which repeat different things at the same time (although sometimes a single verbal form is heard which is accompanied by a non-verbal sound). It is suggested that matching of the auditory input to the particular patterns (or templates) required for perception of syllables or words involves separation of the signal into two fractions. One fraction is matched to the template corresponding to a syllable or word (as modified by a repetition-induced criterion shift). The other fraction corresponds to the residue remaining after subtraction of the components of the auditory input which are used for this match. This residue can appear as a nonlinguistic noise, or it may be matched to a second linguistic template, and thus heard as a different voice repeating some other utterance. The process by which an auditory signal can be synthesized through subtraction of the appropriate components from a louder sound has been called "auditory induction" (Warren, Obusek, & Ackroff, 1972; Warren, 1984). In conjunction with repetition-induced shifts in acoustic criteria defining linguistic templates, auditory induction could facilitate the matching of vowel sequences to syllables and words by permitting the segregation of spectral components corresponding to these modified templates.[1]

It would be of interest to determine the correspondence of individual speech sounds forming the illusory words to the vowels actually present at that time. Preliminary experiments have shown that the mapping of perceptual phonemes to acoustic phonemes can be accomplished, but not through methods that might appear to be the most obvious. Placing an acoustic marker such as a click in one of the vowels doesn't work, since clicks (and other extraneous sounds) are mislocalized in speech (Ladefoged, 1959; Warren & Obusek, 1971). Increasing the intensity of a vowel appreciably and then listening for a corresponding increase in the level of speech sounds in the illusory word doesn't work, since the illusory word usually continues to be heard, and the increased intensity results in hearing the vowel veridically, but as an extraneous sound which cannot be localized in the word. Deleting a vowel and listening for the disappearance of a portion of the illusory word doesn't work, since the illusory word can change to another form. However, a method for phoneme mapping of recorded speech employed in earlier studies (Warren, 1971; Warren & Sherman, 1974) does appear to work quite well. When the repeated sequence of vowels is abruptly terminated, the illusory word or words also stops suddenly, and it is easy to perceive the last speech sound heard. By systematically changing the point of termination of the sequence of vowels, it is possible to map the perceptual phonemes to the acoustic phonemes. Further work employing this procedure in progress.

Summary and Conclusions

Experiment 1 has shown that repeated sequences consisting of different arrangements of the same three vowels can be distinguished either through naming the order of components (for item durations greater than 100 ms) or by recognition of patterns through "temporal compound" identification (for durations from 10 through 100 ms). Perception in a "speech mode" occurred for items from 30 through 100 ms, allowing permuted orders to be discriminated through perception of different verbal organizations for the different arrangements. Nonverbal temporal compounds permitted the discrimination of different arrangements for vowels briefer than 30 ms. Experiments 2a and 2b examined the speech mode of perception further by employing complex repeated sequences consisting of ten 40-ms vowels. The recognition of different arrangements was accomplished readily through verbal mediation even for the minimal changes in order produced by interchanging the position of two contiguous items. The vowel sequences were heard as a single utterance plus a noise, or as two concurrent utterances produced by distinctly different voices.

It was hypothesized that two mechanisms are involved in the illusory perception of words with repeated sequences. The syllabic or lexical templates employed for verbal recognition were temporarily warped into a closer resemblance to the repeated stimulus through repetition-induced "criterion shifts", and matching of the stimulus to the template was then completed by extracting components needed for the match from the auditory input. This perceptual splitting of the stimulus (which also occurs during phonemic restoration) produced a residue which was either perceived as an extraneous sound accompanying the illusory verbal organization or was organized into a second verbal form heard along with the first.

It is of interest that studies with animals other than humans have shown that, although they can discriminate between different arrangements of brief sounds, they fail when the task requires the remembering of sounds for more than a few seconds. As discussed below, this difference between the performance of humans and other animals has suggested how speech perception might have evolved from auditory skills possessed by our prelinguistic ancestors.

Temporal Compounds and the Evolution of Speech. Following a literature survey of studies demonstrating that cats, chinchillas, and monkeys can be taught to recognize not only isolated phonemes, but also monosyllables, the suggestion has been made that the mechanisms employed by humans for speech perception evolved through the elaboration of an ability to recognize overall patterns (or temporal compounds) which we share with other animals (Warren, 1982, 1988). In addition to the animal studies involving sequences of speech sounds, other experiments involving periodic sounds and noises have shown that dolphins (Thompson, 1976) and monkeys (Dewson & Cowey, 1969) can be taught to discriminate between pairs of brief sounds arranged in different orders. However, successful discrimination could be accomplished only when the sequences were brief: when the task required that these animals remember the identity of the first sound for 2 s or more before hearing the second sound, the task became impossible (for further discussion, see Warren, 1982, pp. 137-138). It seems that discrimination between sequences with long separation between items requires a mechanism which is lacking in other animals but available to humans. This mechanism appears to involve verbal encoding so that, for items separated by more than a few seconds, linguistic characterizations (rather than the memory of the sounds themselves) are stored to serve as the basis of discrimination.

For recognition of sequences with brief item durations (such as speech) neither humans nor other animals need identify the order of components or even the components themselves. Only temporal compounds need be recognized. While listeners may be able to name the ordered series of phonemes corresponding to a word, this analytical description does not necessarily imply that the components themselves are perceived. Thus, Brubaker & Warren (1988) demonstrated that listeners can readily learn to name the order of acoustic phonemes corresponding to words that are perceived, even when these words have phonetic transcriptions which do not correspond to the acoustic-phonetic components. They used recycled sequences of three vowels (as in experiment 1). Their subjects first were presented with the two possible arrangements of the vowels at item durations of a few hundred milliseconds (permitting easy identification of order). They then heard these sequences at item durations which were decreased in a regular fashion down to values well below the threshold of 100 ms reported for identification of order with recycled sequences of vowels (Dorman et al., 1975; Thomas et al., 1971). At no time were subjects ever told the actual phonemes or their orders. Through a series of successive generalizations, subjects continued to identify accurately the constituent vowels in their proper orders even though as in the present study, the words heard at brief item durations did not have phonetic transcriptions corresponding to the acoustic phonemes actually present in the stimulus. It was concluded that the perception of syllables and words did not involve a "bottom up" or prior identification of an ordered arrangement of phonetic components. Rather, the identification of the acoustic phonemes and their orders required the mediation of a prior verbal organization.[2]

The recognition of lexical items in connected discourse, of course, consists of more than just the factors described above. Syntactic, semantic and pragmatic rules come into play with lexical aggregates, and these emergent higher-level processes can in turn influence word recognition. However, experiments involving perception of isolated words and phrases (as in the present study) can provide information concerning some of the flexible and opportunistic mechanisms used for the early stages of verbal processing.

References

Bashford, J. A., Jr., & Warren, R. M. (1988). Discrimination of recycled word-length sequences. Journal of the Acoustical Society of America, 84, S141 (Abstract).

Brubaker, B. S., & Warren, R. M. (1988). Learning to identify phonemic orders. Journal of the Acoustical Society of America, 84, S154 (Abstract).

Cole, R. A., & Scott, B. (1973). Perception of temporal order in speech: The role of vowel transitions. Canadian Journal of Psychology, 27, 441-449.

Cullinan, W. L., Erdos, E., Schaefer, R., & Tekieli, M. E. (1977). Perception of temporal order of vowels and consonant-vowel syllables. Journal of Speech and Hearing Research, 20, 742-751.

Dewson, J. H. III, & Cowey, A. (1969). Discrimination of auditory sequences by monkeys. Nature, 222, 695-697.

Diehl, R. L., Kluender, K. R., & Parker, E. M. (1985). Are selective adaptation and contrast effects really distinct? Journal of Experimental Psychology: Human Perception and Performance, 11, 209-220.

Divenyi, P. L. & Hirsh, I. J. (1978). Some figural properties of auditory patterns. Journal of the Acoustical Society of America, 64, 1369-1385.

Dorman, M. F., Cutting, J. E., & Raphael, L. J. (1975). Perception of temporal order in vowel sequences with and without formant transitions. Journal of Experimental Psychology: Human Perception and Performance, 104, 121-129.

Ladefoged, P. (1959). The perception of speech. In National Physical Laboratory Symposium No. 10, Mechanisation of Thought Processes. Her Majesty's Stationery Office, London, 1, 309-417.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. Psychological Review, 74, 431-461.

Nickerson, R. S., & Freeman, B. (1974). Discrimination of the order of the components of repeating tone sequences: Effects of frequency separation and extensive practice. Perception & Psychophysics, 16, 471-477.

Repp, B. H. (1989). Phone restoration. Journal of the Acoustical Society of America, 85, S137 (Abstract).

Skinner, B. F. (1936). The verbal summator and a method for the study of latent speech. Journal of Psychology, 2, 71-107.

Teranishi, R. (1977). Critical rate for identification and information capacity in hearing system. Journal of the Acoustical Society of Japan, 33, 136-143.

Thomas, I. B., Cetti, R. P., & Chase, P. W. (1971). Effect of silent intervals on the perception of temporal order for vowels. Journal of the Acoustical Society of America, 49, 84 (Abstract).

Thomas I. B., & Fitzgibbons, P. J. (1971). Temporal order and perceptual classes. Journal of the Acoustical Society of America, 50, 86-87 (Abstract).

Thomas, I. B., Hill, P. B., Carroll, F. S., & Garcia, B. (1970). Temporal order in the perception of vowels. Journal of the Acoustical Society of America, 48, 1010-1013.

Thompson, R. K. R. (1976). Performance of the bottlenose dolphin (Tursiops truncatus) on delayed auditory sequences and delayed auditory successive discriminations. Doctoral dissertation, University of Hawaii.

Warren, R. M. (1968). Relation of verbal transformations to other perceptual phenomena. Conference Publication No. 42, Institution of Electrical Engineers (London), Supplement No. 1, 1-8.

Warren, R. M. (1971). Identification times for phonemic components of graded complexity and for spelling of speech. Perception & Psychophysics, 9, 358-363.

Warren, R. M. (1972). Perception of temporal order: Special rules for initial and terminal sounds of sequences. Journal of the Acoustical Society of America, 52, 167 (Abstract).

Warren, R. M. (1974). Auditory temporal discrimination by trained listeners. Cognitive Psychology, 6, 237-256.

Warren, R. M. (1982). Auditory perception: A new synthesis. New York: Pergamon Press.

Warren, R. M. (1983). Multiple meanings of "phoneme" (articulatory, acoustic, perceptual, graphemic) and their confusions. In N.J. Lass (ed.), Speech and language: advances in basic research and practice, Vol. 9. New York: Academic Press, pp. 285-311.

Warren, R. M. (1984). Perceptual restoration of obliterated sounds. Psychological Bulletin, 96, 371-383.

Warren, R. M. (1985). Criterion shift rule and perceptual homeostasis. Psychological Review, 92, 574-584.

Warren, R. M. (1988). Perceptual bases for the evolution of speech. In M.E. Landsberg (ed.), The genesis of language. Berlin: Mouton de Gruyter, pp. 101-110.

Warren, R. M., & Ackroff, J. M. (1976). Two types of auditory sequence perception. Perception & Psychophysics, 20, 387-394.

Warren, R. M., & Byrnes, D. L. (1975). Temporal discrimination of recycled tonal sequences: Pattern matching and naming of order by untrained listeners. Perception & Psychophysics, 18, 273-280.

Warren, R. M. & Obusek, C. J. (1971). Speech perception and phonemic restorations. Perception & Psychophysics. 9, 358-362.

Warren, R. M. & Obusek, C. J. (1972). Identification of temporal order within auditory sequences. Perception & Psychophysics, 12, 86-90.

Warren, R. M., Obusek, C. J. & Ackroff, J. M. (1972). Auditory induction: Perceptual synthesis of absent sounds. Science, 176, 1149-1151.

Warren, R. M., Obusek, C. J., Farmer, R. M., & Warren, R. P. (1969). Auditory sequence: Confusion of patterns other than speech or music. Science, 164, 586-587.

Warren, R. M., & Sherman, G. L. (1974). Phonemic restorations based on subsequent context. Perception & Psychophysics. 16, 150-156.

Warren, R. M., & Warren, R. P. (1970). Auditory illusions and confusions. Scientific American, 223 (December), 30-36.

Watson, C. S. (1987). Uncertainty, informational masking, and the capacity of immediate memory. In W. A. Yost & C. S. Watson (eds.), Auditory processing of complex sounds. Hillsdale, New Jersey: Erlbaum, pp. 267-277.

Wickelgren, W. A. (1969). Context-sensitive coding, associative memory, and serial order in (speech) behavior. Psychological Review, 76, 1-15.

Author Notes

This study was supported in part by grants from the National Institutes of Health and the Air Force Office of Scientific Research. We thank Bradley S. Brubaker for his valuable help. Correspondence may be sent to Richard M. Warren, Department of Psychology, University of Wisconsin-Milwaukee, Milwaukee, WI 53201.

Footnotes

[1] Another example of linguistic auditory induction is given by the "phonemic restoration effect" in which contextually appropriate fragments of speech are synthesized from the substrate furnished by a louder sound of appropriate spectral characteristics (for a detailed discussion see Warren, 1984). In keeping with induction theory, Repp (1989) has reported that spectral components corresponding to the restored phoneme are subtracted from an interpolated noise.

[2] It has been suggested that there are four rather different uses of the term "phoneme" (acoustic, articulatory, graphemic, and perceptual), and that confusion has resulted from employing the same term for different entities (Warren, 1983). It was argued that the existence of phonemes as units entering into the perceptual processing of discourse lacks direct experimental support, and that the treatment of perceptual phonemes in the literature is often confounded with acoustic-based phonemes and with articulation-based phonemes.

Table 1

Description of the Stimuli Consisting of Three Recycled Vowels in Experiment 1

Item Duration (ms)	Items per Sequence Burst	Interburst Interval (IBI) in ms	Bursts per Stimulus Set	Stimulus Set Duration in sec
10	301	300	10	32.8
12 *	301	300	10	38.8
30	91	300	10	30.0
100	31	300	8	26.9
300	10	300	8	26.1
1000	10	1000	4	43.0
3000	7	1000	4	87.0
5000	7	1000	4	143.0

*Locked waveforms were used (see text)

Table 2

Accuracy of Same/Different Judgments for Recycled Sequences of Three Vowels in Experiment 1

Stimulus	Responses		Z Scores
Duration of Item (ms)	Number Correct (of 108)	Percent Correct	
10 †	84	78	5.77*
12 ‡	98	91	8.47*
30	90	83	6.92*
100	98	91	8.47*
300	102	94	9.24*
1000	103	95	9.43*
3000	106	98	10.01*
5000	107	99	10.20*

†Judgments were made by a separate groups (see text).
‡These items had locked waveforms (see text).
*p < .0001

Table 3

Numbers of Listeners (out of 32) with Perfect Scores (Correct Identification of Each of the Four 10-Item Vowel Sequences in a Trial) in Experiment 2a

	Practice Trials		**Formal Trials**									
Trial Number	1	2	1	2	3	4	5	6	7	8	9	10
Perfect Scores*	22	27	28	31	30	31	28	31	32	32	27	32

*p < .05 for each perfect score

Table 4

Accuracy and Response Times for ABX Judgments of Recycled Ten-Vowel Sequences in Experiment 2b (A and B Sequences Differed in the *Order of* a Single, Contiguous Pair of Vowels)

		Response Times (seconds)		
Listener	**Number Correct out of 96***	**Median**	**Q_1**	**Q_3**
BB	92	34.5	25.0	51.0
JB	94	50.5	30.0	107.5
JR	94	72.0	41.0	114.0
KR	94	42.0	28.0	68.5

* Accuracy scores for all listeners exceeded chance ($Z \geq 8.98$, $p < .0001$)